Analysis of Metadata and Cough Signal Complexities in Tuberculosis Screening

Sana Ben Mahjouba[1,2][0009−0002−9153−910X], Youssef Ouakrim[1,2][0009−0002−4589−3976], Johannes C. Ayena[1,2][0000−0003−4104−8070], Simon Grandjean Lapierre[3,4][0000−0003−3646−1573], Mihaja Raberahona[5,6][0000−0001−8857−5834], and Neila Mezghani[1,2][0000−0002−5935−4570]

[1] Applied Artificial Intelligence Institute (I2A), TELUQ University, Montréal, Québec, Canada
[2] Open Innovation Laboratory in Health Technologies, Centre de Recherche du Centre Hospitalier de l'Université de Montréal (CRCHUM), Montréal, Québec, Canada
neila.mezghani@teluq.ca
[3] Immunopathology Axis, CRCHUM, Montréal, Québec, Canada
[4] Department of Microbiology, Infectious Diseases and Immunology, Université de Montréal, Montréal, Québec, Canada
[5] CHU Joseph Rasera Befelatanana, Antananarivo, 101, Analamanga, Madagascar
[6] Équipe de Recherche Clinique en Maladies Infectieuses (ERCMI), Madagascar

Abstract. Tuberculosis (TB) remains a global health challenge, particularly in resource-constrained settings where advanced diagnostics are scarce. Recent research has explored the potential of using cough signals, combined with clinical data and machine learning (ML), as a non-invasive screening method for TB. However, the effectiveness of such approaches remains highly dependent on the specific ML model employed, including its architecture, training procedure, and preprocessing pipeline. This study aims to quantify the complexity of both metadata and temporal cough signals to improve classification outcomes and ensure greater model generalizability. Cough recordings from 1105 participants (TB-positive and TB-negative) in the CODA-TB dataset were analyzed across two levels: (1) metadata complexity, assessing demographic and clinical variable heterogeneity; and (2) time series complexity, evaluating raw acoustic waveform irregularity. Complexity metrics, including Shannon entropy, Hurst exponent, and Fisher's discriminant ratio, were computed. Metadata analysis identified fever, weight loss (Cramér's $V > 0.3$) and heart rate (Fisher score > 100) as highly discriminative. Time series analysis revealed TB-specific patterns, with a Hurst exponent of 0.72 ± 0.15 indicating persistent behavior and a Higuchi fractal dimension of 1.71 ± 0.24 reflecting signal irregularity. By integrating complexity analysis, this study uncovers intricate cough patterns and supports the development of robust, non-invasive screening tools for resource-limited settings and highlights their potential for broader clinical applications.

Keywords: Tuberculosis · Cough Acoustics · Complexity Analysis · Machine Learning · Non-invasive Diagnostics

1 Introduction

Tuberculosis (TB) remains a global health challenge, especially in resource-limited settings lacking advanced diagnostics like GeneXpert [16]. Chronic cough, a hallmark of pulmonary TB, has led to acoustic-based diagnostic approaches. However, features such as mel-frequency cepstral coefficients (MFCCs), widely used in audio processing, show variable performance due to environmental noise, recording device differences, and physiological variability, limiting robustness and generalizability [2]. Complexity metrics have gained interest as they capture intrinsic signal properties (temporal dynamics and nonlinear characteristics) that are more robust to noise and acquisition variability [12], providing complementary information that enhances discrimination between TB-positive and TB-negative cases beyond traditional acoustic features.

This study focuses exclusively on active pulmonary TB, excluding latent and extrapulmonary forms. It is designed as a preliminary analysis preceding classification, aiming to identify and characterize complexity-based biomarkers before model training. Unlike previous work investigating only acoustic features [11], we (1) combine metadata complexity analysis with time-series complexity measures of cough sounds; (2) leverage a uniquely diverse dataset from seven countries to ensure biomarker generalizability; and (3) systematically evaluate complexity features to overcome model dependency common in machine learning (ML)-based diagnostics. Clinically, this dual-level analysis reflects TB's multi-faceted nature: cough acoustics indicate pulmonary pathology, while metadata like fever and weight loss capture systemic symptoms. Examining both through complexity theory identifies complementary biomarkers, laying the groundwork for robust TB-positive vs. TB-negative classification.

2 Materials and Methods

2.1 Database

The CODA-TB dataset [9] contains cough audio recordings and associated metadata from 1105 participants (297 TB-positive, 808 TB-negative) across seven countries. This diverse dataset includes (1) metadata such as participant demographics and clinical information, (2) time series data consisting of acoustic cough recordings, and (3) TB status as the target variable indicating TB-negative (0) or TB-positive (1). Together, these components enable a comprehensive analysis for TB detection.

Metadata Metadata include numerical and categorical variables (Tables 1, 2), which provide critical context for the acoustic data and help assess relationships between participant characteristics and cough patterns.

Table 1. Numerical metadata attributes

Variable	Description	Unit	Reference
Age	Participant's age	Years	Nv1
Weight	Participant's weight	Kilograms	Nv2
Height	Participant's height	Centimeters	Nv3
Heart rate	Beats per minute	bpm	Nv4
Temperature	Body temperature	°C	Nv5
Reported cough duration	Duration of cough	Days	Nv6

Table 2. Summary of categorical metadata attributes with reference identifiers

Categorical variables	Description	Values	Reference
Sex	Biological sex	Male / Female	Cv1
TB Prior	Previous TB diagnosis	Yes / No	Cv2
Prior TB type: Pulmonary	Pulmonary TB history	Yes / No	Cv3
Prior TB type: Extrapulmonary	Extrapulmonary TB history	Yes / No	Cv4
Prior TB type: Unknown	Unknown TB type history	Yes / No	Cv5
Hemoptysis	Blood in sputum	Yes / No	Cv6
Weight loss	Recent weight loss (last 30 days)	Yes / No	Cv7
Smoking last week	Recent smoking	Yes / No	Cv8
Fever	Recent fever (last 30 days)	Yes / No	Cv9
Night sweats	Recent night sweats (last 30 days)	Yes / No	Cv10
TB status (Target)	TB diagnosis	1: Yes / 0: No	Cv11

Time Series Data: Acoustic Recordings Cough is a key symptom of pulmonary TB, but symptom-based screening alone often lacks accuracy. Advances in digital health enable acoustic analysis to improve early detection [9]. Cough sounds were recorded using the Hyfe Research App on Android [1] smartphones, which detects cough events via a prediction score threshold (≥ 0.85). Participants were instructed to cough at least three times, with spontaneous coughs also recorded. This time series data enables extraction of acoustic features to distinguish TB-positive from TB-negative coughs using machine learning.

2.2 Complexity Metrics

Complexity analysis quantifies the structure, variability, and discriminative power of both metadata and cough signal time series, aiding feature selection to improve TB screening model interpretability and generalizability [3].

Metadata Complexity Metadata complexity influences classification. Categorical variables were numerically encoded after assessing association with TB status via *Cramér's V* [10], while numerical features were evaluated using complexity metrics that quantify class separability. Let $D = x_1, \ldots, x_n$ be the dataset with features $(f_1, \ldots, f_p)$ and classes c_1 (TB-positive) or c_2 (TB-negative). The applied metrics are summarized in Table 3.

Table 3. Metadata Complexity Measures

Measure	Definition	Interpretation
Fisher Discriminant Ratio (F1) [4]	$F1_i = \frac{(\mu_{i1}-\mu_{i2})^2}{\sigma_{i1}^2+\sigma_{i2}^2}$	Higher values indicate better class separation
Volume of Overlap (F2) [7]	Ratio of overlapping feature value ranges between classes	Values near 0 indicate strong separation; values near 1 indicate high overlap
Intra/Inter-class Distance Ratio (N2) [7]	Ratio of average within-class to between-class distances	Lower values imply better clustering within classes
Shannon Entropy (F4) [15]	$F4_i = -\sum_j p_j \log_2 p_j$	Higher entropy indicates more variability; lower entropy suggests redundancy

Time Series Complexity To capture dynamic and structural cough signal properties, multiple time series complexity measures were applied (Table 4), offering complementary insights for TB detection.

Parameters used in Table 4: Sample Entropy $m = 2$, tolerance $r = 0.2\times$ std; Higuchi max segment length $k_{max} = 10$; Rényi order $\alpha = 2$; Shannon and Rényi computed with 50 histogram bins or KDE bandwidth 0.5; Correlation Dimension embedding $m = 2$, scale range empirical.

Table 4. Time Series Complexity Measures

Measure	Mathematical Definition	Interpretation
Shannon Entropy [15]	$H = -\sum_{i=1}^{n} p_i \log_2 p_i$	Measures signal unpredictability; higher values indicate more randomness
Rényi Entropy [14]	$H_\alpha = \frac{1}{1-\alpha} \log_2 \left(\sum p_i^\alpha\right)$	Generalized entropy sensitive to event frequency; parameter α controls sensitivity
Hurst Exponent [8]	$(R/S)_n = c \cdot n^H$	Assesses long-term memory; $H = 0.5$ indicates randomness, $H > 0.5$ persistence, $H < 0.5$ mean reversion
Higuchi Fractal Dimension [6]	$L(k) \propto k^{-D}$	Quantifies self-similarity; higher D indicates more irregularity
Sample Entropy [13]	$\text{SampEn} = -\ln \frac{A}{B}$	Measures pattern predictability; lower values indicate regularity, higher values more randomness
Correlation Dimension [5]	$D_2 = \lim_{r\to 0} \frac{\log C(r)}{\log r}$	Estimates attractor dimensionality; higher values imply more complex dynamics

3 Results and Discussion

This section reports results on metadata and complexity metrics distinguishing TB-positive from TB-negative coughs, highlighting key findings and limitations.

3.1 Metadata Complexity

The complexity metrics for numerical metadata variables were analyzed to assess their discriminative power. Table 5 presents the Fisher discriminant ratio (F1), overlap volume (F2), intra/inter-class distance ratio (N2), and Shannon entropy (F4) for each variable.

The highest F1 scores, indicating strong class separation, were observed for heart rate (Nv4, 124.16) and temperature (Nv5, 82.65), while reported cough duration (Nv6, 5.28) showed the lowest, suggesting limited discriminative ability. F2 values, measuring class overlap, were lowest for weight (0.31), heart rate (0.40), and temperature (0.41), indicating minimal overlap and good classification potential. N2, reflecting intra- vs. inter-class distances, was lowest for reported cough duration (0.08) and age (0.15), but heart rate (0.36) and temperature (0.30) combined favorable N2 with high F1, reinforcing their relevance. Shannon entropy (F4) varied, with heart rate showing high variability (5.79) and temperature lower (1.16) but still useful.

Overall, heart rate and temperature emerge as the most informative numerical variables due to their high F1 scores, low F2 values, and favorable N2 ratios.

The categorical variables were analyzed using *Cramér's V* metric to quantify their association with the TB status. As shown in Figure 1, the features fever (Cv9), weight loss (Cv7), and night sweats (Cv10) demonstrated the strongest relationships, indicating that these symptoms are significantly more frequent among TB-positive cases.

In summary, the complexity analysis identified key metadata variables relevant for TB screening. Among categorical features, fever, weight loss, and night sweats exhibited the strongest association with TB status (high Cramér's V, Figure 1). For numerical features, heart rate, temperature, and weight demonstrated high Fisher scores (F1), low overlap (F2), and good class separability

Table 5. Complexity metrics for numerical metadata variables

Variable	F1 (Fisher)	F2 (Overlap)	N2 (Intra/Inter)	Shannon Entropy
Nv1	19.24	0.49	0.15	5.74
Nv2	79.40	0.31	0.29	5.55
Nv3	22.68	0.45	0.16	5.10
Nv4	**124.16**	**0.40**	0.36	**5.79**
Nv5	**82.65**	0.41	**0.30**	1.16
Nv6	5.28	0.49	0.08	3.97

Note: Variables Nv1 to Nv6 correspond to numerical metadata attributes described in Table 1.

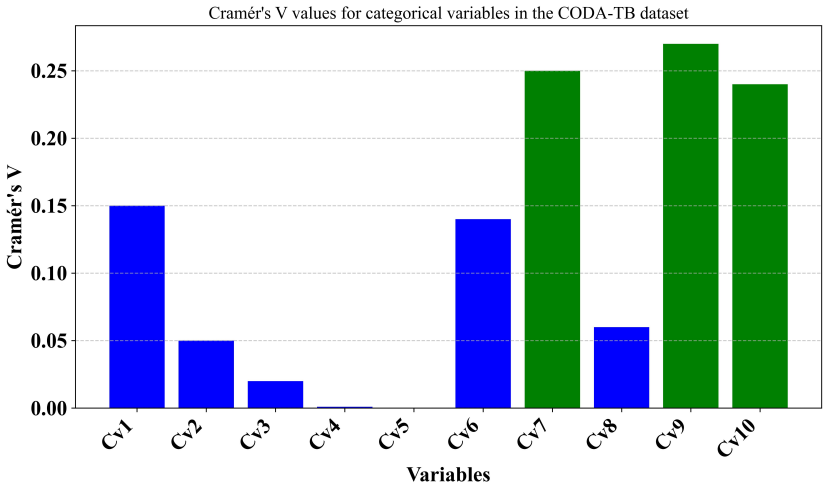


Fig. 1. Cramér's V value for categorical metadata
Note: Categorical variables Cv1 to Cv10 correspond to metadata attributes (Sex, TB Prior, Prior TB type: Pulmonary, Prior TB type: Extrapulmonary, Prior TB type: Unknown, Hemoptysis, Weight loss, Smoking last week, Fever, Night sweats) as defined in Table 2.

(N2) (Table 5). These results highlight the importance of combining clinical symptoms and physiological measures with acoustic features to improve early TB detection.

Table 6. Cough signal complexity metrics

Metric	Mean ± Std	Range
Shannon Entropy	3.64 ± 0.67	$[0, +\infty)$
Rényi Entropy	2.85 ± 0.69	$[0, +\infty)$
Hurst Exponent	0.72 ± 0.15	$[0, 1]$
Higuchi Fractal Dimension	1.71 ± 0.24	$[1, 2]$
Sample Entropy	0.25 ± 0.19	$[0, +\infty)$
Correlation Dimension	0.95 ± 0.38	$[0, m]$

3.2 Time Series Complexity

Table 6 summarizes the complexity metrics computed from cough audio signals, providing insights into signal unpredictability, regularity, and fractal complexity that help differentiate TB-positive and TB-negative coughs.

Entropies (Shannon, Rényi, Sample) quantify signal disorder and unpredictability. The Hurst exponent indicates long-term memory and persistence. Fractal dimensions (Higuchi and correlation) capture geometric irregularity and nonlinear dynamics.

Analysis reveals TB-associated cough signatures: moderate entropy suggests notable unpredictability, a Hurst exponent above 0.5 implies persistent cough patterns, and fractal dimensions reflect moderate complexity. These metrics show promise as discriminative features for non-invasive TB detection.

3.3 Limitations

This study shows the promise of complexity-based features for TB screening but has limitations. Smartphone cough recordings suffer from noise and variability, needing future denoising. The cross-sectional design prevents analysis of cough changes over treatment. No classification models were trained, so diagnostic accuracy wasn't assessed. Future work will compare complexity features with standard ones, include classification, and explore multimodal fusion. Despite CODA-TB's diversity, clinical validation is required before practical use.

4 Conclusion

This study presents a dual-level complexity analysis combining clinical metadata and cough acoustic features to distinguish TB-positive from TB-negative cases. Key discriminative variables like fever, weight loss, heart rate, and acoustic metrics such as the Hurst exponent and Higuchi fractal dimension were identified on a large, diverse dataset. Although classification was not investigated, our current findings form a strong foundation for future predictive models. The focus on interpretable, robust features supports development of scalable, non-invasive TB screening tools for resource-limited settings. Future work will incorporate these features into classification models and validate them clinically.

Acknowledgments. This research was supported by the Canada Research Chair on Biomedical Data Mining (950-231214). The datasets were provided by Drs. Cattamanchi and Grandjean Lapierre in collaboration with international research partners and funded by the U.S. NIH the Patrick J. McGovern Foundation and Global Health Labs. The authors declare no conflicts of interest.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. Hyfe research app. https://www.hyfe.ai/, accessed August 2025
2. Botha, G.H.R., Theron, P., Ackermann, H., Herbst, C.A.: Automatic adventitious respiratory sound analysis: A systematic review. PLOS ONE **13**(5), e0197926 (2018). https://doi.org/10.1371/journal.pone.0197926
3. Dou, Q., Wei, Z., Zhang, Y., Wu, J.: A review of complexity-based data mining methods for healthcare applications. Artificial Intelligence in Medicine **107**, 101–882 (2020)
4. Fisher, R.A.: The use of multiple measurements in taxonomic problems. Annals of Eugenics **7**, 179–188 (1936)
5. Grassberger, P., Procaccia, I.: Measuring the strangeness of strange attractors. Physica D: Nonlinear Phenomena **9**(1–2), 189–208 (1983). https://doi.org/10.1016/0167-2789(83)90298-1
6. Higuchi, T.: Approach to an irregular time series on the basis of the fractal theory. Physica D: Nonlinear Phenomena **31**(2), 277–283 (1988). https://doi.org/10.1016/0167-2789(88)90081-4
7. Ho, T.K., Basu, M.: Complexity measures of supervised classification problems. IEEE Transactions on Pattern Analysis and Machine Intelligence **24**(3), 289–300 (2002)
8. Hurst, H.: Long-term storage capacity of reservoirs. Transactions of the American Society of Civil Engineers **116**, 770–799 (1951). https://doi.org/10.1061/TACEAT.0006515
9. Khobragade, P., Lodha, P., Nagrale, D., Sharma, S., Sable, P.: Detection of tuberculosis using cough audio analysis: a deep learning approach with capsule networks. Discover Artificial Intelligence **4**(1), 22 (2024). https://doi.org/10.1007/s44163-024-00179-4
10. Meyer, D., Zeileis, A., Hornik, K.: The cramér-v and other measures of association: An r perspective. Journal of Statistical Software **76**(1), 1–27 (2017), https://www.jstatsoft.org/article/view/v076i01
11. Rajasekar, S.J.S., Balaraman, A.R., Balaraman, D.V., Ali, S.M., Narasimhan, K., Krishnasamy, N., Perumal, V.: Detection of tuberculosis using cough audio analysis: a deep learning approach with capsule networks. Discover Artificial Intelligence **4**(1), 77 (2024). https://doi.org/10.1007/s44163-024-00179-4
12. Raubitzek, S., Neubauer, T.: Combining measures of signal complexity and machine learning for time series analysis: A review. Entropy **23**(12), 1672 (2021). https://doi.org/10.3390/e23121672
13. Richman, J.S., Moorman, J.R.: Physiological time series analysis using approximate entropy and sample entropy. American Journal of Physiology-Heart and Circulatory Physiology **278**(6), H2039–H2049 (2000). https://doi.org/10.1152/ajpheart.2000.278.6.H2039
14. Rényi, A.: On measures of entropy and information. In: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability. pp. 547–561 (1961), https://projecteuclid.org/euclid.bsmsp/1200512181
15. Shannon, C.E.: A mathematical theory of communication. Bell System Technical Journal **27**(3), 379–423 (1948). https://doi.org/10.1002/j.1538-7305.1948.tb01338.x
16. World Health Organization: Global tuberculosis report 2024. https://www.who.int/teams/global-tuberculosis-programme/tb-reports/global-tuberculosis-report-2024 (2024), accessed May 2025