

Complexity-Pursuit Structural Descriptors for Lung Ultrasound Video Analysis

Marlon Malheiros¹, Thaisse Dias-Paes¹, Ronaldo Zampolo¹ and Antonio Pereira¹

Abstract—Automated lung ultrasound (LUS) analysis is usually built on static image appearance, so the diagnostic value of video dynamics is rarely quantified. This paper addresses that gap with a matched representation-ablation experiment: descriptor families are fixed before any model is fitted and compared on identical videos, splits, and candidate classifiers, so that only the representation varies. Forty-six public LUS videos spanning COVID-19, healthy, bacterial pneumonia, and viral pneumonia were reduced to one descriptor vector each through region-of-interest extraction, analytic-envelope transformation, principal component analysis, and Complexity Pursuit (CP) decomposition. Twenty appearance, six direct-motion, and sixteen CP-derived structural descriptors formed eight predefined families, evaluated by nested repeated-stratified 5×5 cross-validation. In COVID-19-versus-non-COVID-19 classification the combined motion-structural family reached ROC-AUC 0.910 against 0.749 for appearance-only descriptors; a paired out-of-bag bootstrap placed the mean difference at 0.186 (95% interval -0.179 to 0.573). Four-class macro-F1 was 0.543 against 0.386. Perturbation tests indicate the ordering reflects acquisition robustness: crop shift degraded appearance descriptors to correlation 0.857, while motion and structural descriptors held at 0.999. The gain is consistent in direction but imprecise at this cohort size. The study contributes an auditable low-data reference point for how much video dynamics add over static appearance.

Keywords—Lung ultrasound, Complexity Pursuit, representation ablation, feature families, low-data classification, COVID-19.

I. INTRODUCTION

Lung ultrasound (LUS) is widely used for bedside assessment of pneumonia because it is portable, inexpensive, radiation-free, and suitable for repeated monitoring. Recent clinical reviews support its value in pneumonia assessment [1], [2]. Automated LUS analysis remains challenging, however, because public video datasets are often small, heterogeneous, and assembled from multiple acquisition sources. Individual clips may also contain scanner-interface elements, variable probe types, and different acquisition conventions.

Against that background, most automated LUS studies emphasize image appearance or deep visual representations, typically through deep-learning classification, segmentation, or severity-scoring pipelines rather than explicit motion-structural modeling [3], [4]. Recent systems use individual frames, video clips, or complete examinations as inputs [5], [6], [7]. These methods establish the feasibility of automated LUS analysis, but because the representation and the classifier

are optimized jointly, they do not isolate how much of the performance is attributable to video dynamics as opposed to static appearance. Separating the two requires that the representation be varied while everything else is held fixed.

That is the design adopted here. Descriptor families are defined *a priori* and are never selected from data; each family is then passed through an identical evaluation protocol, so any difference in performance is attributable to the descriptors alone. Two of the three families are conventional: static appearance summaries and direct motion summaries. The third is derived from Complexity Pursuit (CP), an unsupervised projection-pursuit method that decomposes a preprocessed video into temporal modes and paired spatial maps [8]. To our knowledge, CP has not previously been applied to lung ultrasound video, and we introduce it here as a general-purpose decomposition for this modality: it is unsupervised, it requires no annotation beyond the video itself, and it yields a linked spatial-temporal pair from which a wide range of descriptors can be derived. The present study exercises one such derivation and evaluates it for classification.

The objective of this study is therefore to determine, under a matched evaluation protocol, whether a video-level representation combining direct motion summaries with CP-derived structural descriptors discriminates better than static appearance descriptors, and to characterize how far that advantage transports across acquisition conditions. The remainder of the paper is organized as follows. Section II describes the cohort, the descriptor construction, the matched evaluation protocol, and the reproducibility settings. Section III reports the family comparison, a fixed-model reference, and the stability and acquisition-sensitivity analyses, each followed by its interpretation. Section IV closes with the main findings and their limits.

II. MATERIALS AND METHODS

The study proceeds in three stages, and the rest of this section follows them in order. First, a fixed video-level manifest is defined and each video is reduced to a single descriptor vector containing appearance, direct-motion, and CP-derived structural summaries (Sections II-A–II-C). Second, the descriptors are partitioned into eight predefined families and compared under one nested cross-validation protocol shared by all families (Section II-D). Third, the selected representation is subjected to uncertainty, stability, perturbation, and transport analyses, and all settings needed to reproduce the results are stated (Section II-E). Throughout, the unit of measurement and of validation is the video: the main task is binary COVID-19-versus-non-COVID-19 classification, and a complementary

¹Marlon Malheiros, Thaisse Dias-Paes, Ronaldo Zampolo and Antonio Pereira are with the Signal Processing Laboratory, Institute of Technology, Federal University of Pará, Belém, PA, Brazil. Corresponding author: Marlon Malheiros, e-mail: marlon.malheiros1@gmail.com.

four-class task separates COVID-19, healthy, bacterial pneumonia, and viral pneumonia.

A. Cohort and Analysis Unit

The source data were drawn from a public COVID-19 LUS repository containing videos and images from multiple acquisition sources [9]. A fixed video-level manifest was defined before feature extraction. Eligible samples were LUS videos with diagnostic labels mappable to COVID-19, healthy, bacterial pneumonia, or viral pneumonia classes; readable video files; sufficient temporal content; successful preprocessing and envelope processing; and probe metadata when available. Still images, unreadable clips, ambiguous labels, preprocessing failures, and entries outside the four target classes were excluded.

Applying these criteria left 46 videos: 8 COVID-19, 17 healthy, 16 bacterial pneumonia, and 5 viral pneumonia. Probe metadata indicated 37 convex and 9 linear videos across 12 source domains, where a source domain denotes the public acquisition source or repository grouping. The videos were associated with 42 examinations, of which three contributed multiple clips. Because adjacent LUS frames are highly correlated, all evaluation used one aggregated row per video; frame-level splitting was never used, as it would place near-duplicate frames on both sides of a split and inflate the resulting estimates.

B. Video Preprocessing and CP Decomposition

All three descriptor families are computed from a common preprocessed representation, so that the comparison in Section II-D is not confounded by differences in preprocessing. Each video was converted to grayscale, temporally subsampled with a fixed stride of 2, and spatially downsampled by a factor of 0.5. The region of interest (ROI) corresponded to the visible ultrasound imaging field, excluding scanner-interface elements, borders, and inactive background. It was estimated from the temporal median frame using Gaussian smoothing (5×5 kernel), Otsu thresholding, morphological closing with a 5×5 elliptical kernel, and bounding-box cropping; the full frame was retained when the mask occupied less than 2% of the frame area. The result is an ROI stack of T frames with $P = HW$ pixels, sketched in Fig. 1.

For each pixel p the grayscale value across frames defines a temporal signal $x_p(t)$, converted to a Hilbert-envelope representation

$$A_p(t) = |x_p(t) + i\mathcal{H}\{x_p(t)\}|, \quad (1)$$

where $\mathcal{H}\{\cdot\}$ denotes the Hilbert transform. Stacking these trajectories gives an envelope matrix $\mathbf{A} \in \mathbb{R}^{P \times T}$, in which each column is one vectorized ROI frame and each row is the envelope trajectory of one ROI pixel. The envelope was used rather than the raw intensity because it emphasizes local temporal amplitude modulation, which is the quantity the decomposition is meant to organize.

Before CP, $\mathbf{A}^\top$ was projected onto a compact principal component analysis (PCA) subspace, treating frames as observations and ROI pixels as variables, yielding scores $\mathbf{Z} \in \mathbb{R}^{T \times d}$ with $d \leq \min(50, T-1, P)$. The scores were whitened before

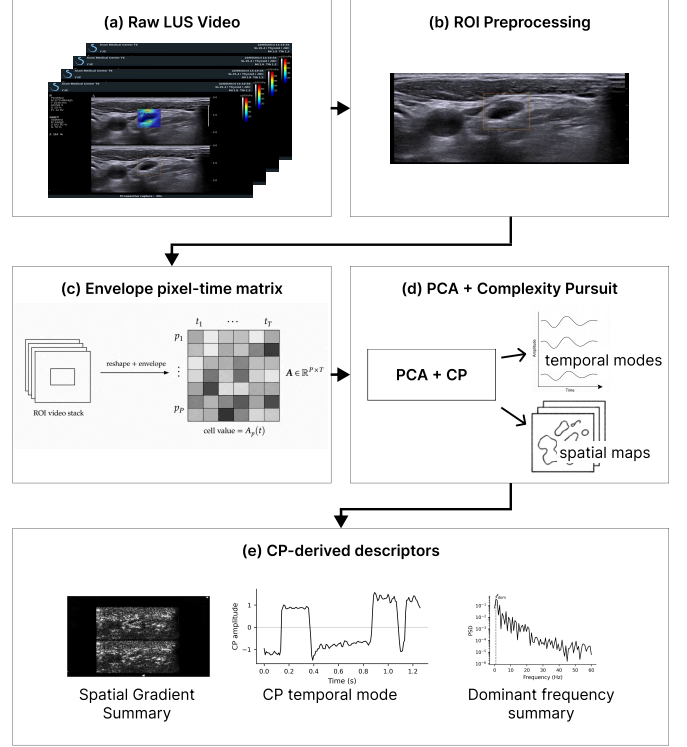


Fig. 1: Workflow of the CP-derived structural representation. (a) Raw exported LUS video frames are used as input. (b) ROI preprocessing crops the visible ultrasound imaging field after removing interface elements and inactive background. (c) The ROI stack is reshaped into an envelope pixel-time matrix, where rows correspond to ROI pixel trajectories and columns correspond to time frames. (d) PCA whitening and Complexity Pursuit decompose the envelope matrix into CP temporal modes and corresponding spatial maps. (e) CP outputs are summarized as per-video descriptors using spatial-gradient summaries of CP maps and dominant-frequency summaries of CP temporal modes.

CP. Writing $\mathbf{z}(t)$ for the whitened representation of frame t , CP estimates projection vectors $\mathbf{w}_k$ favouring non-Gaussian, temporally correlated, low-complexity trajectories [8], giving temporal modes

$$s_k(t) = \mathbf{w}_k^\top \mathbf{z}(t). \quad (2)$$

Up to five modes were retained per video, estimated by normalized gradient updates for at most 500 iterations (tolerance 10^{-6}), combining a sign-based term with a one-frame-lag term weighted by $\alpha = 0.1$, with sequential orthogonalization. The decomposition ran independently within each video and used no diagnostic labels and no information from other videos, so it cannot leak label information into the subsequent evaluation. Projection weights were mapped back through the fitted PCA to reconstruct spatial maps $M_k(u, v)$, so each component yields a linked pair: a temporal mode describing how a latent dynamic pattern evolves, and a spatial map describing where it is expressed. Both members of the pair are summarized in Section II-C; no property of the maps beyond those summaries enters the classifier.

C. Descriptor Construction

Each video was reduced to the 42 descriptors listed in Table I. Appearance and direct-motion descriptors follow conventional definitions. The CP-derived descriptors require

TABLE I: Construction of the per-video classifier descriptors. Feature families are formed by concatenating these groups: appearance-only (20), motion-only (6), structural-only (16, the last three rows), and their combinations up to all-features (42).

Source	n	Per-video summaries
Appearance	20	ROI intensity distribution, entropy, Laplacian and gradient variation, frame variability, regional statistics, and asymmetry.
Direct motion	6	Mean and standard deviation of inter-frame absolute-difference energy, temporal variance, and optical-flow magnitude.
CP spatial maps	9	Mean and maximum spatial-gradient magnitude; regional mean, standard deviation, coefficient of variation, focality, and left-right, center-edge, and top-bottom asymmetry.
CP modes	5	Dominant-frequency mean, standard deviation, and maximum; CP-source standard-deviation and absolute-amplitude summaries.
CP context	2	Retained CP-mode count and retained-frame count.

two summaries. Spatially, for component k we computed the spatial-gradient magnitude

$$S_k(u, v) = \sqrt{(G_x * M_k)^2(u, v) + (G_y * M_k)^2(u, v)}, \quad (3)$$

where G_x and G_y are Sobel operators and $*$ denotes two-dimensional convolution; these maps quantify local spatial variation in the CP mode maps. Temporally, the power spectral density of each mode $s_k(t)$ was estimated with Welch's method [10] at the effective sampling rate after stride-2 subsampling, and the dominant frequency was taken as the highest local spectral peak, with a spectral-centroid fallback when no peak was detected. Summaries were then aggregated across the retained modes, so that the classifier operated on exactly one descriptor vector per video.

D. Matched Family Comparison

The comparison rests on one constraint: the families were fixed before any model was fitted and were evaluated as whole descriptor groups, never as automatically selected feature subsets. Eight families were formed from the three groups in Table I: the three single-group families (appearance-only, 20 features; motion-only, 6; structural-only, 16), the three pairwise combinations (appearance-motion, 26; appearance-structural, 36; motion-structural, 22), the full concatenation (all-features, 42), and an all-plus-probe family (43) adding a single binary variable indicating linear versus convex probe type. The motion-structural family is the one of primary interest, as it contains video dynamics with no appearance descriptors at all; the all-plus-probe family tests whether an explicit acquisition variable changes the ranking.

Every family was then evaluated by nested repeated-stratified 5×5 cross-validation, on the same videos and the same splits. In each outer-training fold a shuffled three-fold inner loop selected one of four prespecified pipelines, listed with their settings in Section II-E. Because model selection is repeated inside each training fold and never sees the outer test videos, the outer estimate remains unbiased with respect to that selection. Two further analyses accompany the family comparison. A paired out-of-bag bootstrap resampled the

TABLE II: Fixed evaluation and classifier settings.

Component	Fixed setting
Gaussian naive Bayes	Variance smoothing 10^{-9} .
Logistic regression	L2 penalty; $C = 1$; L-BFGS optimizer; tolerance 10^{-4} ; 4,000 maximum iterations; balanced class weights.
Random forest	200 trees; maximum depth 6; square-root feature sampling; Gini splitting; balanced class weights.
Gradient boosting	100 estimators; logistic loss; learning rate 0.1; depth 3; Friedman mean-squared-error criterion.
Nested selection	Repeated-stratified outer 5×5 cross-validation; shuffled three-fold inner loop; mean ROC-AUC for binary classification (COVID-19 positive) and mean macro-F1 for four-class classification; ties resolved by the listed pipeline order.
Fixed-model analysis	Standardized, class-balanced logistic regression with pooled leave-one-video-out predictions; scaling fitted on training videos only.
Randomization and software	Seed 42 for randomized splits, bootstrap resamples, and tree-based fits; Python 3.10.11; scikit-learn 1.7.2; NumPy, SciPy, and OpenCV.

video-level rows with replacement, each family independently repeating the inner selection step before evaluation on its held-out videos, giving a distribution for the *difference* between two families rather than for each family separately. Separately, a single standardized, class-balanced logistic-regression pipeline was prespecified for the motion-structural family and evaluated with pooled leave-one-video-out predictions; this provides one transparent threshold-based reference point, complementing the nested comparison, which deliberately averages over candidate models. Scaling and fitting were repeated on training videos only before application to each held-out video.

E. Reproducibility Settings

All candidate configurations were fixed before evaluation; no hyperparameter search was performed. Table II records the classifier, validation, and implementation settings required to reproduce the evaluation. The fixed manifest, derived tables, scripts, estimator settings, and random seed are retained with the project and will be released with the code.

Robustness was assessed with bootstrap-over-video validation using out-of-bag estimation, temporal-window consistency analysis, preprocessing perturbation tests (brightness, crop shift, temporal resampling), leave-one-source-domain-out validation, and a probe-transfer analysis in which the model is trained on one probe type and tested on the other.

III. RESULTS AND DISCUSSION

A. Matched Feature-Family Comparison

The first question is whether the representation matters at all once the protocol is held fixed. Table III shows that it does, and by a wide margin relative to the spread among the families themselves.

Against appearance-only descriptors, the motion-structural family achieved a 0.161 higher binary ROC-AUC point estimate (0.910 versus 0.749). In four-class classification the same contrast gave macro-F1 0.543 versus 0.386 and macro one-versus-rest AUC 0.799 versus 0.715. The paired out-of-bag

TABLE III: Matched feature-family comparison for binary COVID-19-versus-non-COVID-19 and four-class diagnostic classification.

Family	Binary		Multiclass	
	ROC-AUC	Bal. acc.	Macro-F1	Macro AUC
Motion + structural	0.910	0.843	0.543	0.799
Motion only	0.890	0.745	0.410	0.784
Structural only	0.836	0.680	0.482	0.739
Appearance only	0.749	0.660	0.386	0.715
Appearance + structural	0.831	0.631	0.475	0.744
Appearance + motion	0.800	0.659	0.470	0.771
All features	0.846	0.726	0.514	0.785
All + probe	0.842	0.709	0.518	0.782

TABLE IV: Fixed-model binary performance using motion-structural descriptors and pooled leave-one-video-out predictions.

Metric	Value
Accuracy	0.848
Balanced accuracy	0.809
F1-score	0.632
Sensitivity	0.750
Specificity	0.868
ROC-AUC	0.924
PR-AUC	0.779
Brier score	0.091

bootstrap placed the mean binary difference at 0.186, with a 95% interval of -0.179 to 0.573 over 286 valid resamples, each resample repeating inner model selection separately for the two families.

Two things follow. The direction of the effect is consistent: the motion-structural family is first in all four columns of Table III, the bootstrap mean is positive, and the combined family also exceeds each of its own components, improving binary ROC-AUC by 0.020 over motion-only and 0.074 over structural-only, and macro-F1 by 0.133 and 0.061 respectively. The magnitude, however, is not established: the bootstrap interval crosses zero, which is the expected consequence of resampling 46 videos of which only 8 are COVID-19. The defensible reading is that video dynamics carry information that appearance descriptors do not, not that the improvement equals 0.161 in a new cohort. It is also worth noting that adding appearance descriptors to the motion-structural family (the all-features row) *lowers* binary ROC-AUC from 0.910 to 0.846, and that the probe indicator changes almost nothing; at this sample size, the extra 20 appearance dimensions cost more in estimation variance than they contribute in signal.

B. Fixed-Model Binary Performance

The nested comparison averages over four candidate models, which is appropriate for ranking representations but leaves the operating behaviour of any single deployed model unspecified. Table IV supplies that reference using the prespecified logistic-regression pipeline with pooled leave-one-video-out predictions.

At the default 0.5 threshold the model identified 6 of 8 COVID-19 videos and 33 of 38 non-COVID-19 videos, and bootstrap refitting gave mean ROC-AUC 0.929 (95% interval 0.699–1.000), PR-AUC 0.824 (0.333–1.000), and balanced

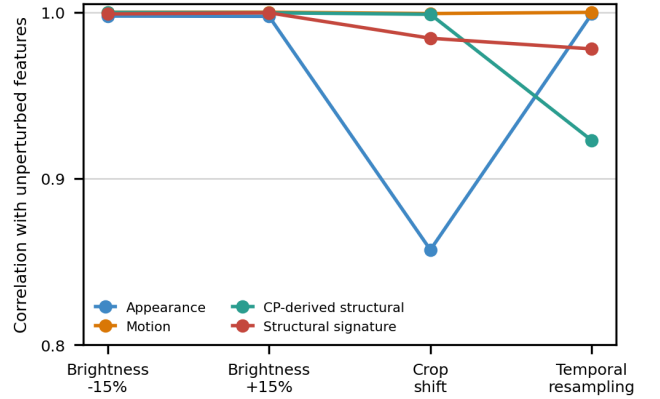


Fig. 2: Mean correlation between unperturbed and perturbed feature vectors across the video cohort. Appearance features were most sensitive to crop shift; temporal resampling affected individual CP-derived structural features more than the aggregate structural signature.

accuracy 0.837 (0.500–1.000). The gap between ROC-AUC 0.924 and F1-score 0.632 is the informative detail here: ranking quality is high while threshold performance is mediocre, because two missed positives out of eight cost a great deal of sensitivity. Any operating threshold estimated from a cohort this size should be regarded as provisional.

C. Descriptor Stability and Acquisition Sensitivity

A representation that ranks well may still be unusable if it is unstable within a clip or dependent on the scanner. We therefore separated descriptor reliability from acquisition dependence.

On reliability, structural signatures extracted from neighbouring temporal windows of the same clip had mean inter-window correlation 0.9976 (95% interval 0.9729–0.99999; $n = 34$), so the descriptors are not an artifact of which segment of a clip is sampled. Under controlled perturbations (Fig. 2), brightness changes had negligible effect on all families (mean correlations ≥ 0.997). Crop shift, however, separated the families sharply: appearance-feature correlation fell to 0.857, whereas motion, structural, and structural-signature correlations remained 0.999, 0.999, and 0.984. Temporal resampling reduced individual structural-feature correlation to 0.923 while the aggregate structural signature retained 0.978.

This perturbation pattern offers a mechanism for the ranking in Table III. Appearance descriptors summarize absolute ROI content and are consequently tied to how the field of view was framed, which varies across the 12 source domains; motion and structural descriptors summarize relative change and are largely invariant to it. The advantage of video dynamics in this cohort is thus at least partly an advantage in acquisition robustness, not solely in pathological specificity.

That interpretation is confirmed by the transport analyses, which also mark the limit of the approach. Leave-one-source-domain-out validation yielded multiclass accuracy 0.448, macro-F1 0.404, and macro one-versus-rest AUC 0.688 — well below the within-cohort figures. Probe transfer was asymmetric: convex-to-linear binary ranking reached ROC-AUC 0.929 yet had zero sensitivity at the default threshold,

meaning the score ordering survived the domain change while its calibration did not; linear-to-convex four-class transfer reached macro-F1 0.174. Acquisition variation, rather than descriptor instability, is therefore the principal constraint on using this representation outside the present cohort.

D. Limitations

Two limitations bound these conclusions. First, cohort size: 46 videos, 8 of them COVID-19 and 5 viral pneumonia, is small enough that every interval reported above is wide, and the four-class results in particular should be read as indicative. Second, no deep-learning baseline is reported. A convolutional or transformer video model has orders of magnitude more parameters than this cohort has videos, so a comparison on these 46 clips would measure regularization and pretraining choices rather than representational merit, and would not be interpretable as evidence either way; establishing competitiveness against modern learned representations requires a substantially larger and more standardized cohort, and is deferred to that setting.

IV. CONCLUSION

This study asked whether an explicit video-dynamics representation discriminates better than static appearance in lung ultrasound classification when nothing but the representation varies. Forty-six public videos were reduced to 42 video-level descriptors through ROI extraction, Hilbert-envelope transformation, PCA whitening, and Complexity Pursuit decomposition, grouped into eight predefined families, and compared by nested 5×5 cross-validation over a fixed bank of four classifiers. Beyond the discrimination results, concatenating appearance descriptors onto the motion-structural family lowered binary ROC-AUC from 0.910 to 0.846, and probe transfer preserved score ordering but not calibration, identifying acquisition variation rather than descriptor instability as the binding constraint. The framework is reproducible from a fixed manifest and fixed estimator settings, and provides a low-data reference point against which richer representations can be measured. Future work should extend the comparison to larger multi-scanner cohorts, where a deep-learning baseline becomes meaningful, and investigate acquisition-invariant normalization to recover the calibration lost under probe transfer.

The source videos were obtained from the public COVID-19 Lung Ultrasound Dataset repository [9]. The analysis code, fixed video-level manifest, extracted descriptors, and scripts used to generate the reported tables and figures will be made available in a public GitHub repository upon publication. The authors acknowledge the publicly accessible ultrasound resources that made this analysis possible.

REFERENCES

- [1] D. Desai *et al.*, “Lung ultrasonography accuracy for diagnosis of adult pneumonia: Systematic review and meta-analysis,” *Advances in Respiratory Medicine*, vol. 92, no. 3, pp. 241–253, 2024.
- [2] B. Kok, P. R. Tuinman, and M. E. Haaksma, “Lung ultrasound in pneumonia: a guide for effective implementation,” *Internal and Emergency Medicine*, vol. 20, no. 2, pp. 357–367, 2025.

- [3] D. Chu, A. Liteplo, N. Duggan, A. B. Hutchinson, and H. Shokoohi, “Artificial intelligence in lung ultrasound,” *Current Pulmonology Reports*, vol. 13, pp. 127–134, 2024.
- [4] M. Muñoz, X. Han, J. Camacho, T. Perrone, A. Smargiassi, R. Inchigolo, Y. Tung-Chen, and L. Demi, “Evaluating deep learning approaches for ai-assisted lung ultrasound diagnosis: an international multi-center and multi-scanner study,” *The Ultrasound Journal*, vol. 17, no. 1, p. 45, 2025.
- [5] M. La Salvia *et al.*, “Deep learning and lung ultrasound for COVID-19 pneumonia detection and severity classification,” *Computers in Biology and Medicine*, vol. 136, p. 104742, 2021.
- [6] J. Diaz-Escobar, N. E. Ordóñez-Guillén, S. Villarreal-Reyes, A. Galaviz-Mosqueda, V. Kober, R. Rivera-Rodriguez, J. E. Rizk, and S. H. Contreras-Ortiz, “Deep-learning based detection of COVID-19 using lung ultrasound imagery,” *PLOS ONE*, vol. 16, no. 8, p. e0255886, 2021.
- [7] Z. Li *et al.*, “Knowledge fused latent representation from lung ultrasound examination for COVID-19 pneumonia severity assessment,” *Ultrasonics*, vol. 143, p. 107409, 2024.
- [8] A. Hyvärinen, “Complexity pursuit: Separating interesting components from time series,” *Neural Computation*, vol. 13, no. 4, pp. 883–898, 2001.
- [9] J. Born, N. Wiedemann, M. Cossio, C. Buhre, G. Brändle, K. Leidermann, A. Aujaeyeb, M. Moor, B. Rieck, and K. Borgwardt, “Accelerating detection of lung pathologies with explainable ultrasound image analysis,” *Applied Sciences*, vol. 11, no. 2, p. 672, 2021.
- [10] P. D. Welch, “The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms,” *IEEE Transactions on Audio and Electroacoustics*, vol. 15, no. 2, pp. 70–73, 1967.