SCImago Journal & Country Rank

Journal Rankings | Journal Value | Country Rankings | Viz Tools | About Us | Search

IEEE Signal Processing Letters

United States | Universities and research institutions | Media Ranking

Country
United States

Subject Area and Category
Computer Science
Signal Processing
Engineering
Electrical and Electronic Engineering
Mathematics
Applied Mathematics

Publisher
Institute of Electrical and Electronics Engineers Inc.

SJR 2025
0.783 Q1

H-Index
173

Publication type
Journals

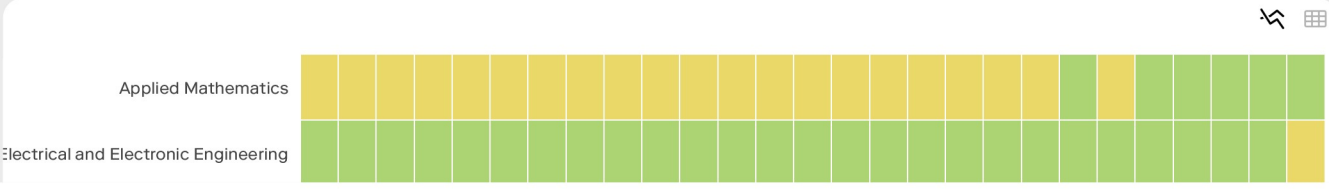
ISSN
10709908, 15582361

Coverage
1994-2026

Information
Home
How to publish in this journal
zjanew@ece.ubc.ca

Scope
The IEEE Signal Processing Letters is a monthly, archival publication designed to provide rapid dissemination of original, cutting-edge ideas and timely, significant contributions in signal, image, speech, language and audio processing.

Quartiles



Category	Year	Quartile
Applied Mathematics	1999	Q2
Applied Mathematics	2000	Q2

Find similar journals

All quartiles | All countries | All subject categories | Clear filters | Download

Only Open Access Journals

1 - Digital Signal Processing: A Review Journal

59%

similarity

2 - Signal, Image and Video Processing

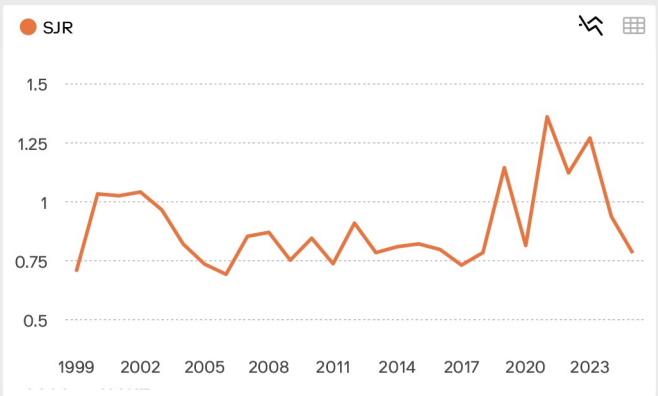
59%

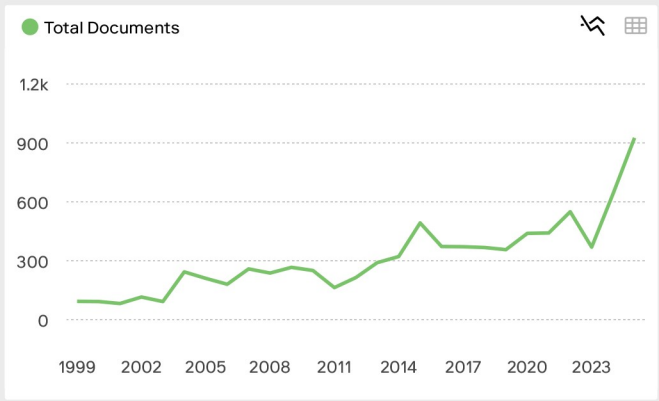
similarity

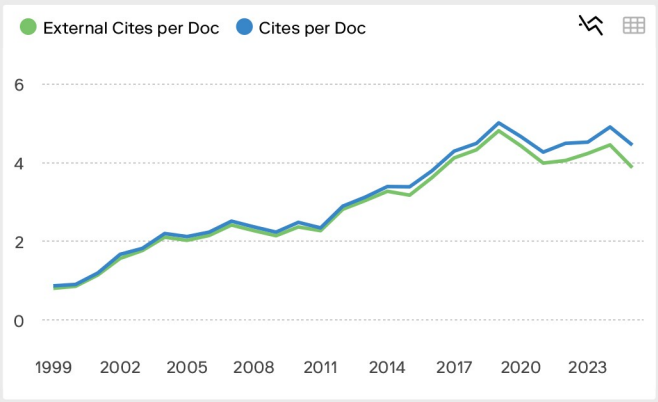
3 - IEEE Open Journal of Signal Processing

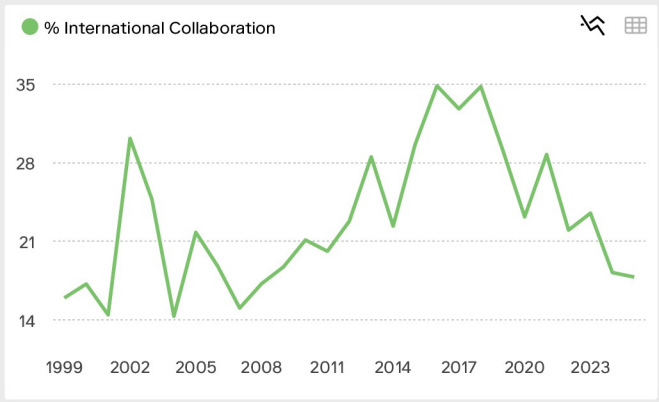
59%

similarity







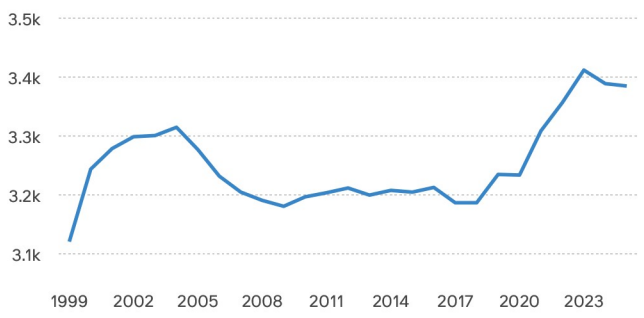


Documents	Year	Value
Uncited documents	1999	164
Uncited documents	2000	166
Uncited documents	2001	132
Uncited documents	2002	126

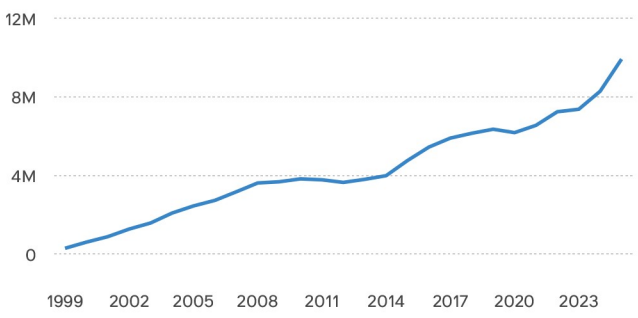
Year	Female Percent
2000	10.96
2001	10.27
2002	12.00
2003	9.60

4 years ago

Estimated APC

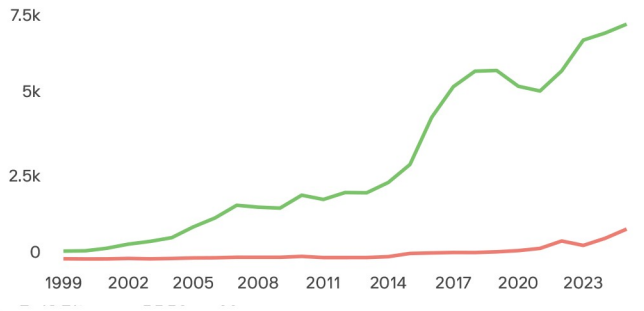


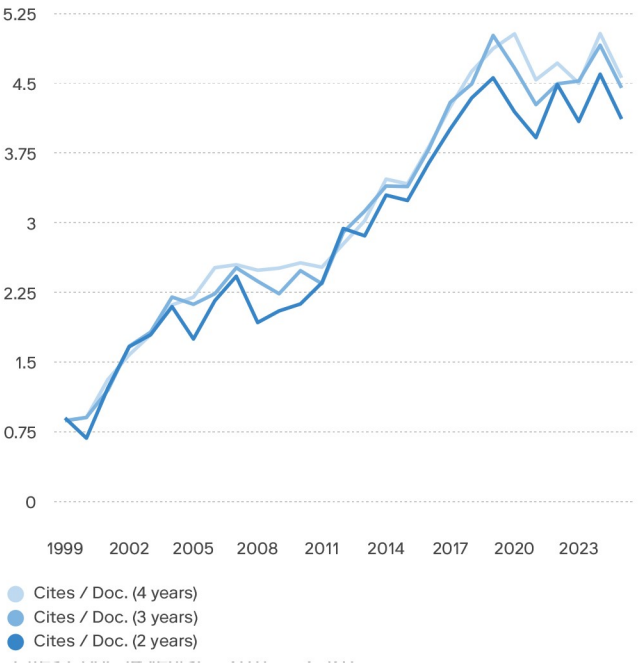
Estimated financial value



Dear Agus, welcome and thanks for your participation! Best Regards, SCImago Team

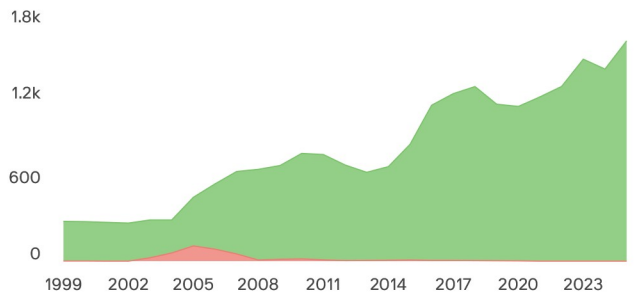
Total Cites Self-Cites





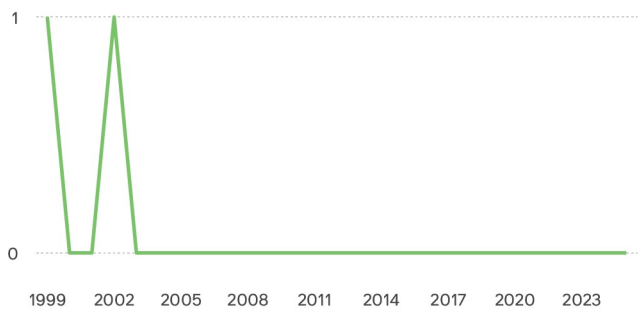
Email

Citable documents Non-citable documents



I'm not a robot

Documents cited by public policy (Overton)

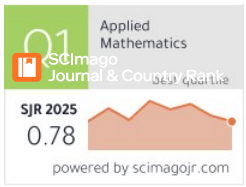


Documents related to SDGs (UN)

Coming soon

IEEE Signal Processing Letters

Show this widget in your



← own website

Just copy the code below and paste within your html code:

<a href="https://www.scimag

SCImago Graphica

Explore, visually communicate and make sense of data with our new data visualization tool.

A Curvature-Controlled Contextual–Perceptual Feature Fusion Framework for ILD Detection from Respiratory Sounds

Publisher: IEEE Cite This

Ayushi Pal ; Udit Satija ; Jimson Mathew ; Hugeng Hugeng ; Choo W. R. Chiong

Abstract:
Interstitial lung disease (ILD) screening from respiratory sounds (RSs) remains challenging due to the subtle acoustic differences between pathological and healthy patter... **Show More**

Metadata

This article has been accepted for publication in IEEE Signal Processing Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LSP.2026.3722466

A Curvature-Controlled Contextual–Perceptual Feature Fusion Framework for ILD Detection from Respiratory Sounds

Ayushi Pal, Udit Satija, Jimson Mathew, Hugeng Hugeng, and Choo W. R. Chiong

***Abstract*—Interstitial lung disease (ILD) screening from respiratory sounds (RSs) remains challenging due to the subtle acoustic differences between pathological and healthy patterns, compounded by limited labeled medical audio data. This paper proposes a curvature-controlled (CC) contextual–perceptual feature fusion framework for automated detection of ILD from RSs. High-level contextual representations extracted using a pre-trained wav2vec 2.0 model are combined with perceptually relevant acoustic features obtained from the encodec neural audio codec. The extracted representations are fused using a geometry-aware fusion strategy operating in a CC latent space. A selective fine-tuning strategy is employed to adapt higher-level acoustic representations to the target task while preserving generalization capability. Experimental results demonstrate that the proposed approach consistently outperforms conventional fusion methods with an accuracy of $86.2 \pm 1.5\%$, highlighting the effectiveness of curvature-aware multimodal representation learning for RS analysis.**

***Index Terms*—Respiratory sounds (RSs), wav2vec 2.0, encodec, curvature-controlled, hyperbolic space, ILD classification.**

I. Introduction

INTERSTITIAL lung disease (ILD) constitutes a diverse spectrum of chronic pulmonary conditions marked by parenchymal inflammation and fibrosis, imposing substantial global health burdens through progressive respiratory failure [1]. While high-resolution computed tomography (HRCT) and invasive biopsies remain gold-standard diagnostics, their resource intensity limits accessibility in resource-constrained settings, underscoring the need for scalable, non-invasive alternatives [2]. Auscultation-based respiratory sound (RS) analysis emerges as a compelling paradigm, harnessing portable acoustic sensors; nevertheless, the nuanced differentiation of adventitious sounds, such as fine crackles and high-pitched wheezes, from normative vesicular patterns challenges traditional feature engineering and Euclidean classifiers [3].

In recent years, many studies have investigated automated RS analysis for detecting respiratory diseases through CT scans [4] or HRCT images [3]. In [5], the authors examined ILD by focusing on the region of interest in HRCT images and obtained features through a refined attention pyramid network that is subsequently complemented by mobileUnetV3 for classification, achieving an accuracy of 99.1%. In [4], the authors implemented a five-layer convolutional neural network (CNN) combined with average pooling to categorize CT images into seven distinct classes with an accuracy of 85.5%. In [6], the authors employed features derived from texture, morphology, and intensity, utilizing a random forest classifier (RFC) to categorize lung tissue patterns in HRCT images, achieving an accuracy of 85.8%. Likewise, the researchers in [3] applied CT scans of patients to examine ILD through an ensemble network made up of InceptionV3, VGG16, and ResNet50, resulting in an accuracy of 82.7%. Recently, some studies [7] examined ILD through RS signals. In [7], the authors introduced a sinc convolutional network for differentiating ILD from healthy cases, achieving an accuracy of 81.25%.

Recent developments in deep learning have greatly enhanced representation learning for analyzing biomedical signals. Convolutional and recurrent neural networks have been extensively utilized on spectrogram-derived RS representations to identify abnormal occurrences like crackles and wheezes [8]. Nonetheless, these models depend on static receptive fields and primarily local temporal relationships, constraining their capacity to grasp long-range contextual details and hierarchical acoustic structures connected to ILD. Additionally, their effectiveness and generalization are frequently limited by the lack of labeled clinical data. Self-supervised learning has become a powerful method for acquiring strong acoustic representations from extensive unlabeled datasets. In [9], the authors used the idea of hyperbolic fusion for speech recognition using pre-trained models. Models like wav2vec 2.0 [10] acquire contextualized representations directly from raw audio waveforms, facilitating the modeling of long-term temporal dependencies without the need for task-specific supervision. Although these embeddings reflect rich high-level context, they might not completely maintain the detailed perceptual traits that are clinically significant for analyzing RSs.

Neural audio codecs, such as encodec [11], deliver compact and perceptually refined latent representations that maintain intricate acoustic details. Although promising, codec-based features have seen limited research in RS classification, and their combination with self-supervised contextual representations is still largely unexamined. Moreover, many current fusion methods depend on Euclidean-space techniques, like feature concatenation, which frequently fall short in adequately

This work is supported by the two projects funded by ANRF, GoI, with diary number ANRF/F/65/2025-2026 & file number CRD/2024/000745 and ANRF/ARGM/2025/000443/TS.

Ayushi Pal and Udit Satija are with the Dept. of Electrical Engineering (EE) at the Indian Institute of Technology Patna (IITP), 801106, Bihar, India.
Jimson Mathew is with the Dept. of CSE at IITP, 801106, Bihar, India.
Hugeng Hugeng is with the Dept. of EE, Universitas Tarumanagara, Jakarta Barat, Indonesia.
Choo W. R. Chiong is with Dept. of Electrical and Computer Engineering, Curtin University Malaysia, Miri, 98009, Sarawak, Malaysia.
(e-mail: (ayushi_2221ee06, udit, jimson)@iitp.ac.in, hugeng@ft.untar.ac.id, raymond.ccw@curtin.edu.my)

This article has been accepted for publication in IEEE Signal Processing Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LSP.2026.3722466

representing intricate interactions among diverse acoustic features.

Motivated by these limitations, this framework proposes a contextual–perceptual feature fusion (CCFF) framework for automated ILD detection from RSs. The proposed approach integrates self-supervised contextual representations extracted using wav2vec 2.0 with perceptually optimized codec-based features obtained from encodec. A geometry-aware fusion mechanism with curvature control is employed to enhance feature interaction and better capture complex acoustic relationships. Additionally, a selective fine-tuning strategy is adopted to adapt higher-level acoustic representations to the target task while preserving the generalization benefits of self-supervised pretraining.

The key objectives of this study are as follows: (i) to investigate the effectiveness of integrating self-supervised contextual embeddings and codec-based perceptual features for RS classification; (ii) to explore a CC, geometry-aware fusion strategy for enhancing feature interaction beyond conventional Euclidean fusion approaches; (iii) to evaluate the role of selective fine-tuning in adapting pre-trained acoustic representations for ILD detection; and (iv) to assess the robustness and generalization capability of the proposed framework across different curvature values. The structure of the paper is as follows: Section II outlines the datasets utilized in our framework, Section III presents the proposed methodology with preprocessing steps and the implementation of pre-trained models for contextual-perceptual FF, Section IV contains the results and discussion regarding the proposed framework, and finally, Section V concludes the entire framework.

II. Dataset Description

This framework employs two publicly accessible RS datasets. The main dataset is the BRACETS [12], which offers RS recordings paired with related electrical impedance tomography (EIT) data from healthy subjects and those with respiratory conditions. Recordings are obtained with a 3M Littmann Electronic Stethoscope 3200 at a sampling rate of 4 kHz from multiple anterior and posterior chest sites in bell, diaphragm, and extended recording modes. This framework considers recordings from healthy subjects and patients with ILD only. The primary BRACETS dataset exhibited substantial class imbalance, comprising 384 recordings from 24 ILD subjects and 176 recordings from 8 healthy subjects. To address this imbalance, all healthy recordings from the KAUH dataset are incorporated into the experimental pool. The KAUH dataset is recorded using the same electronic stethoscope model and sampling frequency, with 5-30 s duration; however, only the healthy recordings are utilized in the present framework. Furthermore, to ensure a more rigorous evaluation and to mitigate concerns related to subject overlap and dataset-source confounding, both segment-wise and subject-level splitting protocols are employed in the framework.

III. Proposed Framework

This proposed framework consists of preprocessing of raw RSs, followed by contextual and perceptual feature extraction and CCFF, and classification of ILD and healthy RSs. A schematic representation of the proposed framework is shown in Fig. 1.

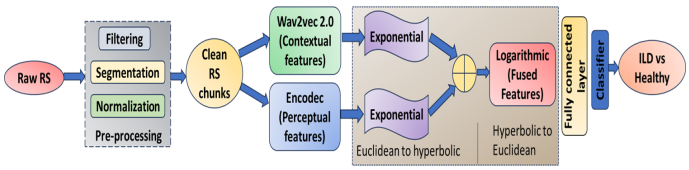


Fig. 1. Block Diagram of the Proposed Framework.

A. Data Preprocessing

The preprocessing phase comprises three consecutive steps: (i) band-pass filtering, (ii) signal segmentation, and (iii) z-score normalization. The raw RS (RRS) recordings are first processed using a band-pass filter with a passband of 10-2000 Hz to preserve the frequency components. The filtered RS signal (F_{RS}) is subsequently segmented into fixed-length windows of 5 s with a 50% overlapping window [13], resulting in a total of 2,247 RS segments.

The segmentation process is mathematically expressed as

$$S_{RS}^{(k)}[n] = F_{RS}[kH + n], \quad 0 \leq n < L, \; k = 0, 1, \ldots, K-1, \tag{1}$$

where $S_{RS}^{(k)}[n]$ denotes the k^{th} segmented RS signal, L is the frame length corresponding to a 5 s segment, $H = L/2$ is the hop size corresponding to a 50% overlap between consecutive segments, and K represents the total number of extracted segments. Following segmentation, each RS segment is normalized using z-score normalization to obtain zero-mean and unit-variance samples, thereby reducing inter-segment variability. The normalized RS segments (N_{RS}) are then provided as input to the wav2vec 2.0 and EnCodec models for contextual and perceptual feature extraction, respectively, for ILD detection.

B. Proposed CCFF Framework for ILD Detection

The proposed framework utilizes N_{RS} for contextual and perceptual feature extraction in parallel.

1) Contextual Feature Extraction Using Wav2vec 2.0: A pretrained wav2vec 2.0 model is utilized to capture high-level contextual information and long-range temporal dependencies found in RSs. Wav2vec 2.0 works directly with raw waveforms and includes a convolutional feature encoder succeeded by a transformer-based contextual encoder [10]. Provided with a N_{RS} chunk, the model generates a series of contextual embeddings, each capturing both local and global sound information. Since the output of wav2vec 2.0 is a sequence of frame-level embeddings, temporal mean pooling is applied across the sequence to obtain a fixed-length representation. This pooled representation, denoted as y_{CFW}, serves as the contextual feature vector and captures global acoustic features associated with ILD related RSs.

2) Perceptual Feature Extraction Using EnCodec: Simultaneously with contextual feature extraction, perceptual acoustic features are obtained through the encodec neural audio codec pretrained model [11]. Encodec acquires concise latent representations tailored for perceptual accuracy and maintains subtle acoustic nuances that might be missed by models focused solely on context. For every N_{RS} chunk, the encodec

encoder produces a multi-channel latent representation across time. To obtain a fixed-dimensional perceptual feature vector, temporal average pooling is applied across the codec output. The resulting feature vector, denoted as y_{PFE}, captures low-level acoustic features such as subtle variations in breath sounds and crackle-like events, which are clinically relevant for ILD detection.

3) CCFF: The contextual feature vector y_{CFW} and perceptual feature vector y_{PFE} represent complementary RS information. Instead of fusing these features using simple concatenation, the proposed framework employs a CC fusion strategy for $c > 0$ to enable richer feature interaction as in [9]. First, both feature vectors are mapped from Euclidean space to a curved latent space using an exponential mapping controlled by a curvature parameter c as:

$$\exp_{\mathbf{c}}(\mathbf{y}) = \tanh\big(\sqrt{c}\|\mathbf{y}\|/2\big)\,\frac{2\mathbf{y}}{\sqrt{c}\|\mathbf{y}\|}, \tag{2}$$

where $\mathbf{u}_{\text{CFW}} = \exp_c(\mathbf{y}_{\text{CFW}})\,, \quad \mathbf{u}_{\text{PFE}} = \exp_c(\mathbf{y}_{\text{PFE}})$ as hyperbolic points for contextual and perceptual feature vectors.

In the curved latent space, FF is performed using a curvature-aware aggregation operator, denoted by $\oplus_c$, which enables non-linear interaction between the u_{CFW} and u_{PFE} representations:

$$\begin{aligned}\mathbf{h}_{\text{fused}} &= \mathbf{u}_{\text{CFW}} \oplus_c \mathbf{u}_{\text{PFE}} \\ &= \frac{(1+2c\langle \mathbf{u}_{\text{CFW}}, \mathbf{u}_{\text{PFE}}\rangle + c\|\mathbf{u}_{\text{PFE}}\|^2)\mathbf{u}_{\text{CFW}} + (1-c\|\mathbf{u}_{\text{CFW}}\|^2)\mathbf{u}_{\text{PFE}}}{1+2c\langle \mathbf{u}_{\text{CFW}}, \mathbf{u}_{\text{PFE}}\rangle + c^2\|\mathbf{u}_{\text{CFW}}\|^2\|\mathbf{u}_{\text{PFE}}\|^2}\end{aligned} \tag{3}$$

and produces the fused hyperbolic feature set.

After fusion, the combined representation is mapped back to euclidean space using a logarithmic mapping:

$$\mathbf{y}_{\text{fused}} = \log_{\mathbf{c}}(\mathbf{h}_{\text{fused}}) = \frac{1}{2\sqrt{c}}\,\text{arctanh}(\sqrt{c}\|\mathbf{h}_{\text{fused}}|)\frac{2\mathbf{h}_{\text{fused}}}{\|\mathbf{h}_{\text{fused}}|}, \tag{4}$$

where $\mathbf{y}_{\text{fused}}$ serves as the input to the Euclidean classifier.

4) Classification and Optimization: The contextual branch initially produces 768-dimensional embeddings, whereas the perceptual branch generates 256-dimensional embeddings. Prior to hyperbolic fusion, the contextual embeddings are transformed into a 256-dimensional representation through a fully connected layer, thereby aligning both feature representations within a common latent space. The aligned contextual and perceptual features are then fused, and the resulting Euclidean fused embedding, y_{fused}, is passed to a fully connected classification layer for binary classification of ILD and healthy RSs. The classifier produces class probabilities using the softmax function and is trained with the cross-entropy loss. To balance generalization and task-specific adaptation, a selective fine-tuning strategy is adopted, in which only the higher-level layers of the contextual feature extractor are fine-tuned while the lower-level layers remain frozen. The model is optimized using the Adam optimizer with a learning rate of 1×10^{-4}, a batch size of 32, and 100 training epochs for different curvature values, with c = 0.01, 0.05, 0.1, 0.2, 0.5.

The curvature parameter c determines the extent of hyperbolic expansion in the fusion manifold: lower values resemble Euclidean geometry for a stable, conservative feature integration, while higher values improve hierarchical separation of pathological patterns yet may lead to metric distortion and optimization difficulties. This systematic analysis determines the curvature that best aligns hyperbolic geometry with the inherent structure of respiratory embeddings, thus enhancing discriminative performance and generalization.

IV. Results and Discussions

The performance of the proposed framework is evaluated using various performance measures such as accuracy (Acc.) [14], recall (Rec.) [14], precision (Prec.) [14], specificity (Spec.) [14], F1-score [14], and area under the receiver operating characteristic curve (AUC-ROC) [15] values.

A. Performance Evaluation

To evaluate the performance of the proposed framework segment-wise, the dataset comprising 2,247 RS segments is partitioned into 70% training ($N_{\text{train}} = 1,573$), 15% validation ($N_{\text{val}} = 337$), and 15% testing ($N_{\text{test}} = 337$) subsets. The segmented RS samples are subsequently used for model training, validation, and testing.

In our proposed framework, two wav2vec 2.0 fine-tuning strategies are systematically investigated:

- Last layer fine-tuning (Last FT): Only the final transformer encoder layer is unfrozen; all preceding layers remain frozen.
- Last three layers fine-tuning (Last Three FT): The final three transformer encoder layers are unfrozen and jointly optimized, increasing adaptation capacity while preserving stable low-level features.

Table I presents the performance of the proposed framework across different positive curvature values for the segment-level analysis, reported as mean $\pm$ standard deviation over multiple independent runs using both the **Last FT** and **Last Three FT** strategies. The proposed framework achieves the highest accuracy of $86.2 \pm 1.5\%$ at $c = 0.01$ using the Last Three FT strategy. The results indicate that positive curvature values, particularly in the range of 0.05 to 0.20, provide the most consistent and effective performance in terms of accuracy, F1-score, and AUC-ROC. Furthermore, the Last Three FT strategy consistently outperforms the Last FT strategy across most curvature settings, demonstrating that fine-tuning additional encoder layers enables better adaptation of the pretrained representations to the ILD classification task. Overall, these findings demonstrate that an appropriate choice of positive curvature, together with deeper fine-tuning of the pretrained encoder layers, leads to improved classification performance and better generalization.

The discriminative capability of the learned feature representations is further illustrated using the t-SNE visualizations [13] shown in Fig. 2. The feature representations before training exhibit considerable overlap between the healthy and ILD classes, indicating limited discriminative capability. In contrast, the proposed CCFF produces more compact and well-separated clusters, demonstrating enhanced class separability. Specifically, Fig. 2(a) illustrates the random feature distribution, whereas Fig. 2(b) presents the learned feature representations after training with the proposed CCFF. Fig. 3(a) presents the confusion matrix corresponding to the best-performing

TABLE I
PERFORMANCE MEASURES OBTAINED USING THE PROPOSED FRAMEWORK.

Curvature (c)	Last FT (All values are in %)						Last Three FT (All values are in %)					
	Acc.	Prec.	Rec.	Spec.	F1	AUC-ROC	Acc.	Prec.	Rec.	Spec.	F1	AUC-ROC
0.01	79.9 ± 1.0	77.6 ± 1.1	81.4 ± 2.2	78.7 ± 1.5	79.4 ± 1.1	89.1 ± 0.7	86.2 ± 1.5	84.0 ± 1.0	85.1 ± 2.9	84.7 ± 1.0	84.9 ± 1.8	93.2 ± 0.7
0.05	82.7 ± 0.6	77.9 ± 1.0	89.0 ± 1.7	77.0 ± 1.7	84.1 ± 1.6	92.6 ± 0.9	84.6 ± 0.6	82.2 ± 1.0	86.5 ± 2.6	82.9 ± 1.6	84.3 ± 0.8	92.8 ± 0.4
0.1	83.9 ± 1.1	77.8 ± 0.4	92.5 ± 2.3	76.0 ± 0.8	84.5 ± 1.3	92.9 ± 0.7	84.8 ± 0.8	82.8 ± 0.4	86.0 ± 2.4	83.5 ± 0.6	84.3 ± 1.0	92.3 ± 0.6
0.2	83.5 ± 0.6	79.6 ± 0.7	87.9 ± 1.1	79.5 ± 1.0	83.5 ± 0.5	92.0 ± 0.3	85.4 ± 1.2	84.3 ± 0.7	85.1 ± 3.3	85.6 ± 1.0	84.7 ± 1.4	92.7 ± 0.7
0.5	80.7 ± 1.8	77.8 ± 4.0	83.7 ± 3.3	78.0 ± 5.6	80.5 ± 1.4	88.7 ± 1.4	83.9 ± 1.8	81.8 ± 2.2	85.2 ± 4.5	82.6 ± 3.1	83.4 ± 2.2	91.5 ± 0.6

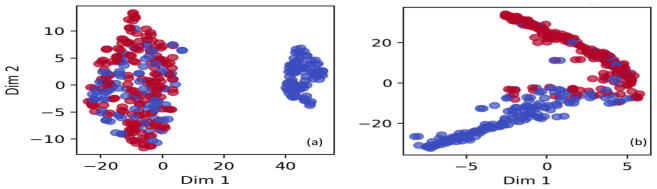


Fig. 2. t-SNE visualization of feature representations: (a) feature distribution before training the proposed framework, and (b) learned feature representations after training with the proposed CCFF, demonstrating improved class separability between healthy and ILD RSs.

curvature ($c = 0.01$), demonstrating high classification accuracy for both healthy and ILD RS segments, with only a small number of misclassifications. The confusion matrix indicates balanced predictive performance across both classes. Furthermore, Figs. 3(b) and 3(c) illustrate the training and validation accuracy and loss curves, respectively. The steadily increasing accuracy together with the decreasing loss demonstrates stable convergence during training with minimal signs of overfitting, supporting the effectiveness of the proposed fine-tuning strategy.

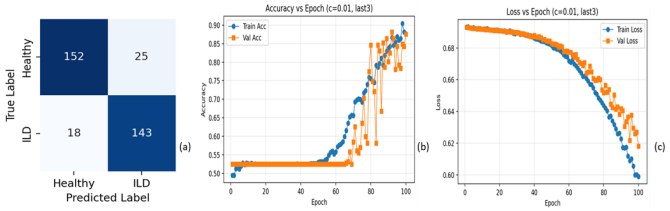


Fig. 3. Performance visualization of the proposed framework at the optimal curvature ($c = 0.01$): (a) confusion matrix, (b) training and validation accuracy curves, and (c) training and validation loss curves.

For the **subject-level ablation study**, 70% subjects (N_{train} = 48) have been kept in the training set, 15% (N_{val} = 10) in validation, and the remaining 15% (N_{test} = 9) have been used as a test set to ensure no data leakage. Table II presents the subject-level performance of the proposed framework across different positive curvature values using the **Last FT** and **Last Three FT** strategies. Among the evaluated curvature values, $c = 0.2$ achieves the best overall performance with the Last Three FT strategy, yielding the highest accuracy, F1-score, and AUC-ROC. Furthermore, the Last Three FT strategy outperforms the Last FT, indicating that fine-tuning the final three encoder layers enhances the robustness of the learned representations and improves subject-level generalization.

The bar chart shown in Fig. 4 presents a comparative analysis of different feature extraction and fusion strategies in terms of Acc., F1-score, Rec., Spec., and AUC-ROC. The results show that fusion-based approaches provide more balanced performance than the individual feature representations, with gated fusion outperforming simple Euclidean fusion. Among

TABLE II
SUBJECT-LEVEL CURVATURE-WISE PERFORMANCE.

Curvature (c)	Last FT (%)						Last Three FT (%)					
	Acc.	Prec.	Rec.	Spec.	F1	AUC-ROC	Acc.	Prec.	Rec.	Spec.	F1	AUC-ROC
0.01	36.7	31.4	84.3	14.6	45.8	58.3	54.9	36.6	58.1	53.4	44.9	57.1
0.05	40.3	32.1	79.0	22.3	45.7	59.1	64.8	45.7	57.6	68.2	51.0	67.8
0.1	36.6	32.4	92.1	10.8	48.0	69.5	57.2	40.7	76.0	48.5	53.0	69.8
0.2	31.0	30.4	91.3	3.0	45.6	48.7	70.4	52.8	62.5	74.0	57.2	73.3
0.5	31.2	31.2	96.9	0.6	47.2	37.6	61.5	43.5	71.2	57.0	54.0	70.7

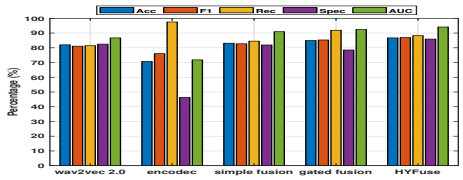


Fig. 4. Comparison of performance measures across different models for ILD detection.

all the compared methods, the proposed CCFF framework achieves the best overall performance, demonstrating the effectiveness of curvature-controlled hyperbolic fusion in learning complementary multimodal representations.

B. Performance Comparison

Table III compares the proposed framework with existing imaging- and RS-based ILD detection methods. Imaging approaches using CT/HRCT [3], [4], and [6] report accuracies between 82.7% and 85.8% but require expensive imaging procedures. Among RS-based methods, [7] achieved 81.25% accuracy using SincNet. In contrast, the proposed CCFF framework achieves $86.2 \pm 1.5\%$ accuracy, $84.9 \pm 1.8\%$ F1-score, and $93.2 \pm 0.7\%$ AUC-ROC using only RS recordings. The improved performance demonstrates the effectiveness of CC contextual–perceptual FF and highlights its potential as a non-invasive, and scalable ILD screening solution.

TABLE III
COMPARATIVE PERFORMANCE ANALYSIS

S. No.	Authors	Data	Method	Results
1	Anthimopoulos et al. [4]	CT images	5-layer CNN architecture	85.5% Acc.
2	Vishraj et al. [6]	HRCT images	Intensity, texture and morphology features with RFC	85.8% Acc., 82.2% F1-score, 83.5% Prec., and 81.7% Rec.
3	Martinez et al. [3]	CT scans	Ensemble network consisting of InceptionV3, VGG16 and ResNet50	82.7% Acc.
4	Arka et al. [7]	RSs	SincNet	81.25% Acc., 78.85% Sens., and 83.33% Spec.
5	Proposed framework	RSs	CCFF	$86.2 \pm 1.5\%$ Acc., $84.0 \pm 1.0\%$ Prec., $85.1 \pm 2.9\%$ Rec., $84.7 \pm 1.0\%$ Spec., $84.9 \pm 1.8\%$ F1-score, $93.2 \pm 0.7\%$ AUC-ROC

V. CONCLUSION

This framework presents a CC contextual–perceptual FF approach for automated ILD detection using RS recordings. By combining self-supervised contextual and perceptually optimized codec-based features, the method effectively captures complementary acoustic information and achieves improved detection performance and robustness using CCFF framework over existing imaging- and sound-based methods.

REFERENCES

[1] B. T. SOCIETY and S. Committee, "The diagnosis, assessment and treatment of diffuse parenchymal lung disease in adults," *Thorax*, vol. 54, no. Suppl 1, p. S1, 1999.

[2] W. D. Travis, U. Costabel, D. M. Hansell, T. E. King Jr, D. A. Lynch, A. G. Nicholson, C. J. Ryerson, J. H. Ryu, M. Selman, A. U. Wells *et al.*, "An official american thoracic society/european respiratory society statement: update of the international multidisciplinary classification of the idiopathic interstitial pneumonias," *American Journal of Respiratory and Critical Care Medicine*, vol. 188, no. 6, pp. 733–748, 2013.

[3] J. B. Martinez and G. Gill, "Comparison of pre-trained vs domain-specific convolutional neural networks for classification of interstitial lung disease," in *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*, 2019, pp. 991–994.

[4] M. Anthimopoulos, S. Christodoulidis, L. Ebner, A. Christe, and S. Mougiakakou, "Lung pattern classification for interstitial lung diseases using a deep convolutional neural network," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1207–1216, 2016.

[5] S. R. Vinta, B. Lakshmi, M. A. Safali, and G. S. C. Kumar, "Segmentation and classification of interstitial lung diseases based on hybrid deep learning network model," *IEEE Access*, vol. 12, pp. 50 444–50 458, 2024.

[6] R. Vishraj, S. Gupta, and S. Singh, "Ecm-iltp: An efficient classification model for categorization of interstitial lung tissue patterns," in *2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, 2021, pp. 481–485.

[7] A. Roy and U. Satija, "Ildnet: A novel deep learning framework for interstitial lung disease identification using respiratory sounds," in *2024 International Conference on Signal Processing and Communications (SPCOM)*, 2024, pp. 1–5.

[8] L. Pham, H. Phan, R. Palaniappan, A. Mertins, and I. McLoughlin, "Cnn-moe based framework for classification of respiratory anomalies and lung disease detection," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 8, pp. 2938–2947, 2021.

[9] O. C. Phukan, Girish, M. M. Akhtar, S. R. Behera, P. B. Reddy, A. B. Buduru, and R. Sharma, "Hyfuse: Aligning heterogeneous speech pre-trained representations in hyperbolic space for speech emotion recognition," 2025. [Online]. Available: https://arxiv.org/abs/2506.03403

[10] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[11] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *Transactions on Machine Learning Research*, 2023.

[12] D. Pessoa, B. M. Rocha, C. Strodthoff, M. Gomes, G. Rodrigues, G. Petmezas, G.-A. Cheimariotis, V. Kilintzis, E. Kaimakamis, N. Maglaveras, A. Marques, I. Frerichs, P. de Carvalho, and R. P. Paiva, "Bracets: Bimodal repository of auscultation coupled with electrical impedance thoracic signals," *Computer Methods and Programs in Biomedicine*, vol. 240, p. 107720, 2023.

[13] A. Pal, A. Roy, and U. Satija, "A unified joint contrastive triplet loss with temporal and frequency signal fusion for diagnosing heart murmurs," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[14] A. Pal and U. Satija, "Leveraging graph node weighted pcg signal propagation for heart valve disorder detection," in *2024 IEEE 21st India Council International Conference (INDICON)*, 2024, pp. 1–6.

[15] P. Adeodato and S. Melo, "A geometric proof of the equivalence between auc_roc and gini index area metrics for binary classifier performance assessment," in *2022 International Joint Conference on Neural Networks (IJCNN)*, 2022, pp. 1–6.