Harnessing Transformer-based Language Models for Interstitial Lung Diseases Diagnosis from Time-frequency Encodings of Lung Sounds

Ayushi Pal[‡], Naseem Babu[†], Udit Satija[‡], Jimson Mathew[†], Hugeng Hugeng[¥], and Choo W. R. Chiong[§]

[‡]Department of Electrical Engineering, Indian Institute of Technology Patna, Patna, Bihar, India; [†]Department of Computer Science and Engineering, Indian Institute of Technology Patna, Patna, Bihar, India; [¥]Department of Electrical Engineering, Universitas Tarumanagara, Jakarta Barat, Indonesia; [§]Department of Electrical and Computer Engineering, Curtin University Malaysia, Miri, Sarawak, Malaysia
Emails: ayushi_2221ee06@iitp.ac.in, udit@iitp.ac.in, naseem_2021cs22@iitp.ac.in, jimson@iitp.ac.in, hugeng@ft.untar.ac.id, raymond.ccw@curtin.edu.my

Abstract—Early and accurate diagnosis of interstitial lung disease (ILD) remains a major clinical challenge due to its wide range of symptoms and overlapping sound patterns with other pulmonary disorders, such as fine crackles and squawks. This paper proposes a transformer-based language model (LM) framework for classifying respiratory sounds associated with ILD using three complementary time-frequency (TF) encodings: constant-Q transform (CQT), mel-spectrograms, and mel-frequency cepstral coefficients (MFCC). Each TF representation is first derived from pre-processed lung sound recordings and then reshaped into token-like sequences suitable for processing by pre-trained transformer backbones, specifically GPT-2, GPT-2-medium, Qwen, and Qwen-chat models. The proposed framework is evaluated on publicly available respiratory sound datasets, BRACETS and KAUH, and performance is reported as average accuracy, recall, precision, and F1-score to provide a comprehensive assessment. The proposed framework underscores the potential of transformer-based LM architectures operating on TF representations of RSs as a promising direction for developing robust, non-invasive, and scalable tools for early ILD screening and for discovering reliable acoustic biomarkers in clinical and real-world applications.

Index Terms—Lung sounds (LSs), interstitial lung disease (ILD), language model (LM), and classification.

I. INTRODUCTION

Interstitial Lung Disease (ILD) encompasses a diverse group of chronic respiratory disorders characterized by progressive scarring and fibrosis of the pulmonary interstitium, leading to impaired gas exchange and reduced lung compliance [1]. ILD has a significant clinical burden, and many of its subtypes have a progressive trajectory that, if left untreated, can result in diminished lung function, a lower quality of life, and early death [2]. Early identification is particularly challenging because symptoms such as exertional dyspnea and dry cough are nonspecific, and access to high-resolution imaging and expert interpretation may be limited in many situations [3]. Abnormal lung sounds (LSs), such as squawks and fine crackles, are significant bedside markers of ILD and are caused by structural and functional changes in the alveolar units and small airways [4]. Traditional diagnostic methods, such as high-resolution computed tomography (HRCT) and pulmonary function tests, though effective, are expensive, time-consuming, and expose patients to radiation [5]. With advances in digital auscultation and signal processing, LSs have emerged as a promising non-invasive modality for computer-aided screening and phenotyping of diffuse parenchymal lung diseases [5].

Over the past few years, numerous studies have explored automated LS analysis for lung disease detection using CT scans [4] or HRCT images [6]. In [7], the authors analyzed ILD within the region of interest (ROI) in HRCT images, extracted features using a refined attention pyramid network (RAPNet), and then used mobileUnetV3 for classification. In [4], the authors used a five-layer convolutional neural network (CNN) followed by average pooling to classify CT images into seven different classes. The authors in [8] used features based on texture, morphology, and intensity, followed by a random forest (RF) classifier to classify lung tissue patterns in HRCT images. Similarly, the authors in [6] used CT scans of patients to analyze ILD using an ensemble network composed of InceptionV3, VGG16, and ResNet50. Recently, a few works, [3], [9], analyzed ILD using LS signals. In [9], the authors proposed a sinc convolutional network for ILD vs healthy classification, resulting in an accuracy of 81.25%.

Recent advancements in deep learning have introduced more flexible and hierarchical models for biomedical sounds, CNNs [4] have demonstrated promise in recognizing respiratory patterns and adventitious sounds such as crackles and wheezes. However, these models rely on fixed receptive fields and sequential recurrence, which may limit their ability to capture the complex temporal–spectral relationships present in ILD signals. Moreover, their interpretability and transferability across datasets remain limited, particularly in low-data or imbalanced settings that are common in clinical studies.

The emergence of transformer architectures and pre-trained LMs [10]–[12] has significantly advanced representation learning across diverse sequential modalities. Models such as GPT [13] and Qwen [14], originally developed for natural language processing tasks, have demonstrated strong capabilities in contextual representation learning and long-range dependency modeling. Recent studies such as TIME-LLM and LLMTime have explored adapting LMs for non-textual temporal data using prompting and sequence-serialization strategies [15], [16], suggesting that these models can be extended beyond textual processing by transforming temporal signals into token-compatible sequential embeddings.

Motivated by these observations, this study investigates the feasibility of employing frozen LMs, including GPT2, GPT2-medium [13], Qwen, and Qwen-chat [14], for the classification of ILD versus healthy lung sounds. Unlike prior prompt-based or autoregressive serialization approaches that primarily focus on forecasting tasks [15], [16], the proposed framework directly processes time-frequency representations of respiratory acoustic signals through structured

This article has been accepted for publication in IEEE Sensors 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/LSENS.2026.3702525

patch-based embeddings and input reprogramming. This enables the model to learn contextual spectro-temporal dependencies without converting the signals into textual prompts or numerical autoregressive token sequences. Furthermore, unlike conventional CNN-based lung sound classification approaches that primarily focus on local feature extraction, the proposed framework leverages transformer-based sequential modeling to capture both local and global dependencies within respiratory acoustic representations.

The key contributions of this work are summarized as follows:

- Introduced a framework that adapts transformer-based LMs (GPT and Qwen families) for LS classification.
- Extracted and encoded time–frequency features like mel-spectrograms, mel-frequency cepstral coefficient (MFCC), and constant q-transform (CQT) representations as structured sequential tokens for model input embedding space.
- Provided a systematic performance comparison between GPT2, GPT2-medium, Qwen, and Qwen-chat architectures, analyzing their ability to generalize and learn from limited LS data.

This method provides the first evidence that pre-trained LMs can be fine-tuned and repurposed for LS analysis. The paper is structured as follows: Section II describes the databases used; Section III presents the proposed methodology, including preprocessing and details of implementing transformer-based LMs; Section IV presents the results and discussion of the proposed framework; and Section V concludes the framework.

II. DATABASE DESCRIPTION

This section provides details on the publicly available databases used in this proposed framework. First one is BRACETS database [17], which consists of LS recordings with associated electrical impedance tomography (EIT) of healthy as well as diseased patients. These recordings were collected using a 3M Littmann electronic 3200 stethoscope at a sampling rate of 4000 Hz from different anterior and posterior regions in three modes: bell, diaphragm, and extended. In our proposed framework, we primarily focus on the ILD and healthy subjects. It is highly imbalanced, with 24 ILD and 8 healthy subjects, comprising of 384 and 176 LS recordings respectively, ranging from 15 to 20 seconds. To mitigate the imbalance issue, we used another database, i.e., KAUH [18], which comprises 35 healthy subjects and various lung diseases, including asthma and pleural effusion. The LS recordings are done in a similar way using a Littmann 3200 stethoscope, having the same sampling rate, ranging from 5 to 30 seconds. Each LS recording is further preprocessed for ILD vs healthy classification.

III. PROPOSED FRAMEWORK

This section includes the details of the proposed framework for classifying ILD and healthy LSs using transformer-based LMs. The pipeline consists of three major stages: (i) preprocessing of raw LS signals, (ii) time–frequency feature extraction, such as MFCC, mel-spectrograms, and CQT, and (iii) description of LMs for LS classification. The block diagram of the proposed framework is shown in Fig. 1.

A. Preprocessing

The preprocessing step consists of three stages: (i) filtering, (ii) segmentation, and (iii) normalization. The raw LS signals (RLS) are first filtered with a band-pass filter from 10 to 2000 Hz. The filtered

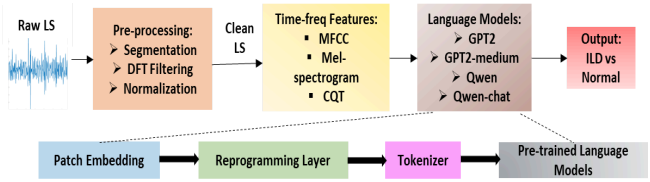

Fig. 1. Block Diagram of Proposed Framework.

LS signals are then segmented into 5 second fixed-length frames with a 50% overlap. Let $\mathrm{LS_f}[n]$ denote the filtered lung sound signal, let L denote the frame length in samples corresponding to a duration of 5 seconds, and let O denote the overlap length in samples, where $O = \lfloor 0.5L \rfloor$. The m-th segmented frame is defined as

$$\mathrm{LS}_m[n] = \mathrm{LS_f}(m(L-O)+n), \quad n = 0, 1, \ldots, L-1, \quad m = 0, 1, \ldots, M-1, \quad (1)$$

where $\mathrm{LS}_m[n]$ represents the m-th segmented LS frame and M denotes the total number of frames extracted from the filtered LS signal. Using this segmentation procedure, a total of 2247 5-second frames are obtained. These chunks are then normalized using z-score normalization to the range [-1, 1]. These normalized LS signals LS_n are then converted to the time-frequency (TF) domain, such as the mel-spectrogram, mel-frequency cepstral coefficients (MFCC), and the constant-Q transform (CQT), which are used as inputs to LMs.

B. Time-frequency Feature Encodings

Three different TF features, such as Mel-spectrogram, MFCC, and CQT extracted from LS_n.

1) Mel-spectrogram: Mel-spectrogram [2] transforms NLS into a spectrogram with frequency bins aligned to the Mel scale. It is generated by performing a short-time Fourier transform (STFT) and then applying a Mel filter bank, producing a 2D representation with rows corresponding to Mel-frequency bands and columns indicating time. This characteristic captures both time-related changes and spectral elements, rendering it very efficient for LSs classification.

2) MFCC: Mel-frequency cepstral coefficient (MFCC) [19] is obtained from the mel-spectrogram through the application of a discrete cosine transform (DCT) on the logarithm of filter bank energies. This method removes correlations among the coefficients and condenses the spectral data into a reduced number of features, usually ranging from 13 to 20 coefficients, that illustrate the signal's spectral envelope. In analyzing LSs, MFCCs help recognize patterns associated with various conditions by focusing on the most significant spectral features.

3) CQT: CQT [20] representation featuring a logarithmically spaced frequency axis, offering greater frequency resolution at low frequencies and reduced resolution at high frequencies. The CQT is obtained by convolving the signal with a series of filters whose bandwidths are proportional to their center frequencies. The resultant representation effectively captures the harmonic architecture and time progression of respiratory signals.

C. Language Models (LMs) for LSs Classification

This study presents a transformer-based LM framework that utilizes three complementary TF representations of LSs, such as mel-spectrograms, MFCC, and CQT, for the classification of ILD and healthy LS signals. The representations are saved as numpy tensors, resized to a standardized structure with channels aligned to frequency

This article has been accepted for publication in IEEE Sensors 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/LSENS.2026.3702525

coefficients while maintaining the temporal dimension, enabling uniform processing, and then divided into training, validation, and test subsets to preserve class balance. A common backbone architecture is used, comprising a patch embedding followed by a reprogramming layer, and frozen, pre-trained LMs such as GPT2, GPT2-medium, Qwen, and Qwen-chat as the sequence model. Time-frequency representations are tensors of shape (N, C, T), where C denotes the number of frequency bins (channels) and T represents the number of temporal frames. To convert these representations into transformer-compatible tokens, a 1D convolutional patch embedding layer is used to perform temporal tokenization. Specifically, a convolution with kernel size p and stride s is applied along the time dimension, treating frequency bins as input channels. This operation produces a sequence of patch tokens according to $(B, C, T) \rightarrow (B, T_{\text{new}}, d_{model})$, where B denotes the batch size, d_{model} denotes the embedding dimensionality of each token, and $T_{\text{new}} = \left\lfloor \frac{T-p}{s} \right\rfloor + 1$. Then, each token (d_{model}) is projected into the pre-trained LMs embedding space (d_{lm}) via a learnable reprogramming layer $(d_{model} \rightarrow d_{lm})$. The resulting tensor of shape $(B, T_{\text{new}}, d_{lm})$ is then input to the frozen LMs using the `inputs_embeds` interface, ensuring compatibility with the original embedding space while preserving the structure of the pre-trained model. The final hidden states are combined via adaptive average pooling, followed by a fully connected classification layer that converts the pooled representation into ILD versus healthy classes. Every model is trained for each TF representation (mel-spectrogram, MFCC, CQT), enabling analysis across them. Training utilizes 100 epochs with an AdamW optimizer, a StepLR learning rate scheduler, and cross-entropy loss with label smoothing to reduce overfitting and enhance generalization. For each representation, the evaluation is performed over 10 independent runs using subject-wise data partitioning, and the final results are reported as the average across all runs. These results enable comparisons among mel-spectrogram, MFCC, and CQT-based LLM models for ILD-affected LS classification.

IV. RESULTS AND DISCUSSION

This section presents the experimental results and analysis of the proposed transformer-based LM framework for ILD versus healthy LS classification using TF encodings of LS signals.

A. Performance Metrics and Evaluation

We evaluate the effectiveness of our proposed framework by employing various performance metrics, including average accuracy (Avg. Acc.), average precision (Avg. Pre.), average recall (Avg. Rec.), and average F1-score (Avg. F1) on the test set. To evaluate the performance of the proposed framework, the dataset is partitioned subject-wise into an 80% training set, a validation set 10%, and a test set 10%. This ensures that no segments from the same individual appear across training, validation, or test sets, thereby preventing overlap and providing a clinically realistic assessment of generalization performance.

TABLE 1 outlines the classification results of various transformer-based LMs employing mel-spectrogram, MFCC, and CQT TF representations for distinguishing ILD from healthy cases. Among all models, GPT2-medium delivers the highest overall performance for mel-spectrogram and MFCC features, achieving average accuracies of $84.31 \pm 1.42\%$ and $83.20 \pm 2.28\%$, respectively, and consistently good precision, recall, and F1-scores. This demonstrates robust discriminative ability and equitable learning for both classes. Conversely, GPT2 and Qwen exhibit lower recall metrics, suggesting a higher likelihood of misclassification, particularly for ILD samples. Regarding CQT features, Qwen-chat surpasses the competing models, reaching the top average accuracy of $79.33 \pm 1.57\%$ and F1-score of $77.54 \pm 3.43\%$. This underscores the effectiveness of Qwen-chat in modeling logarithmic frequency representations, in which GPT-based models exhibit diminished robustness. These observations emphasize that model–feature alignment is crucial, and that no single model is universally optimal across all TF representations.

The confusion matrices, as shown in Fig. 2, also reinforce these conclusions by demonstrating prediction behavior for each class. GPT2-medium shows fewer false positives and false negatives across mel-spectrogram and MFCC TF representations, suggesting enhanced class separability and more reliable ILD detection. In contrast, for CQT features, Qwen-chat demonstrates increased true positive rates for both ILD and healthy categories, validating its enhanced performance noted in TABLE 1.

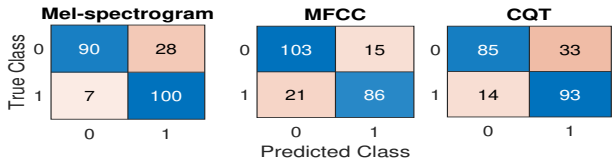

Fig. 2. Confusion matrices for ILD versus healthy classification using (a) GPT2-medium for mel-spectrogram, (b) GPT2-medium for MFCC, and (c) Qwen-chat for CQT TF representation.

Moreover, the t-SNE visualizations shown in Fig. 3 provide qualitative insights into feature distinguishability. Embeddings produced by GPT2-medium for mel-spectrogram and MFCC TF features show more distinct cluster separation between ILD and healthy classes, with little overlap. Conversely, CQT-derived embeddings exhibit tighter, better-separated clusters when analyzed with Qwen-chat, confirming their quantitative advantage. The integrated quantitative and qualitative findings highlight the efficacy of the proposed framework and underscore the importance of selecting appropriate TF features and transformer-based LM architectures for ILD classification.

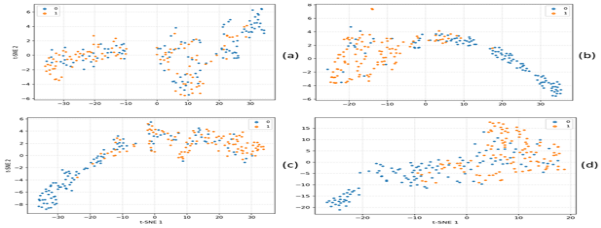

Fig. 3. t-SNE visualization of test set of ILD and healthy embeddings, (a) test set before training, (b) and (c) after training using GPT2-medium for mel-spectrogram and MFCC, and (d) Qwen-chat for CQT.

Fig. 3 (a) shows the random scattering of test samples before training the proposed framework. Fig. 3 (b) and (c) show the clusters formed after training of GPT2-medium for mel-spectrograms and MFCC, and Fig. 3 (d) shows the clusters formed for Qwen-chat for CQT TF encodings.

The ablation study outlined in TABLE 2 investigates the performance of patching, reprogramming and LLM integration in the proposed framework for ILD vs healthy classification. The results show that the overall performance of the reprogrammed LM variant is outperforming on all the performance metrics, indicating that the reprogramming and LM-based representation learning are

TABLE 1. Performance Measures (in %) Obtained Using Proposed Framework for ILD vs Healthy Classification.

Model	Mel-spectrogram				MFCC				CQT			
	Avg. Acc.	Avg. Pre.	Avg. Rec.	Avg. F1	Avg. Acc.	Avg. Pre.	Avg. Rec.	Avg. F1	Avg. Acc.	Avg. Pre.	Avg. Rec.	Avg. F1
GPT2	84.13 ± 0.70	84.45 ± 0.66	84.13 ± 0.70	84.13 ± 0.70	83.11 ± 2.11	83.57 ± 2.00	83.11 ± 2.11	83.10 ± 2.12	79.02 ± 1.49	79.35 ± 1.49	79.02 ± 1.49	78.99 ± 1.50
GPT2 medium	**84.31 ± 1.42**	84.49 ± 1.44	84.31 ± 1.42	84.30 ± 1.42	**83.20 ± 2.28**	84.35 ± 2.03	83.20 ± 2.28	83.11 ± 2.36	78.27 ± 2.11	78.51 ± 2.12	78.27 ± 2.11	78.22 ± 2.17
Qwen	82.84 ± 1.96	83.11 ± 2.03	82.84 ± 1.96	82.83 ± 1.95	78.89 ± 2.38	79.41 ± 2.48	78.89 ± 2.38	78.84 ± 2.40	77.73 ± 3.24	78.29 ± 2.92	77.73 ± 3.24	77.54 ± 3.43
Qwen chat	**83.42 ± 2.01**	**83.77 ± 1.94**	**83.42 ± 2.01**	**83.41 ± 2.02**	**82.58 ± 0.72**	82.88 ± 0.80	82.58 ± 0.72	82.57 ± 0.72	**79.33 ± 1.57**	79.48 ± 1.45	79.33 ± 1.57	79.29 ± 1.63

TABLE 2. Ablation Analysis of the Proposed Framework Using Different Model Variants.

Model Variant	Patch	Reprogramming	LM	Avg. Acc.	Avg. Pre.	Avg. Rec.	Avg. F1
Classifier	✓	✗	✗	72.00%	73.23%	72.00%	71.85%
Frozen LM	✓	✗	✗	81.78%	81.78%	81.78%	81.76%
Reprogrammed LM	✓	✓	✓	84.31%	84.49%	84.13%	84.30%

beneficial. The performance of the classifier only and frozen LMs are comparatively lower, underscoring the critical role of adaptive reprogramming in enhancing the classification ability of proposed framework.

B. Performance Comparison

This section compares the proposed framework with the existing methods of detecting ILD. TABLE 3 shows that most of the previous reported works focus on using CNN- and ensemble-based architectures with the use of CT or HRCT images as the diagnostic modality. These imaging techniques provide detailed information of the lungs in terms of its structure and its anatomy, thus aiding the classification process. The proposed framework, on the other hand, takes as input modality the LS signal, which is a more challenging problem because lung sounds have limited, non-stationary, low cost, radiation free, and highly variable acoustic characteristics. To the best of our knowledge, there has been only one study Arka et.al. [9] that explored the use of LS signals for ILD classification with an accuracy of 81.25%. The proposed framework based on transformer-based LMs and TF representations shows superior performance with these features, even when using only the acoustic information. These results show the effectiveness and potential of the proposed non-invasive ILD classification framework, especially with the mel-spectrogram features of $84.31 \pm 1.42\%$ and the MFCC features of $83.20 \pm 2.28\%$.

TABLE 3. 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. [8]	HRCT images	Features based on intensity, texture, and morphology followed by RF	85.8% Acc, 82.2%, F1-score, 83.5% Prec., and 81.7 % Rec.
3	Martinez et al. [6]	CT scans	Ensemble network consisting of InceptionV3, VGG16 and ResNet50	82.7% Acc.
4	Arka et al. [9]	Lung sounds	SincNet	81.25% Acc., 78.85% Sens., and 83.33% Spec.
5	Proposed framework	Lung sounds	TF features with transformer-based LMs	$84.31 \pm 1.42\%$ Acc. for mel-spectrogram, $83.20 \pm 2.28\%$ for MFCC, and $79.33 \pm 1.57\%$ for CQT

V. CONCLUSION

This study presents a novel framework for the automated classification of ILD versus healthy LSs using TF signal encodings processed through transformer-based LMs. By transforming spectro-temporal features into structured sequential tokens, we successfully repurposed transformer-based LM architectures such as GPT2, GPT2-medium, Qwen, and Qwen-chat for biomedical sound interpretation. The results demonstrate that these models, despite being originally designed for natural language understanding, can effectively capture temporal and spectral dependencies within LS data. Among all tested architectures, GPT2-medium achieved the highest accuracy of approximately $84.31 \pm 1.42\%$ for mel-spectrogram, outperforming conventional deep learning models.

VI. ACKNOWLEDGEMENT

This work is supported by the two projects funded by ANRF, GoI, i.e, ANRF/F/65/2025-2026, and ANRF/ARGM/2025/000443/TS.

REFERENCES

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

[2] A. Roy and U. Satija, "A novel melspectrogram snippet representation learning framework for severity detection of chronic obstructive pulmonary diseases," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–11, 2023.

[3] 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.

[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] D. Vaghela, D. A. Jasani, R. V. Patel, D. N. Patel, M. Patel, and M. Maheshwari, "Role of hrct in the diagnosis of interstitial lung diseases: A cross-sectional study of radiological patterns and demographic correlates," *European Journal of Cardiovascular Medicine*, vol. 15, no. 6, pp. 284–288, Jun 2025.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] S. Zhou, Z. Xu, M. Zhang *et al.*, "Large language models for disease diagnosis: a scoping review," *npj Artificial Intelligence*, vol. 1, p. 9, 2025.

[11] J. Zhou, H. Li, S. Chen *et al.*, "Large language models in biomedicine and healthcare," *npj Artificial Intelligence*, vol. 1, p. 44, 2025.

[12] S. Madan, M. Lentzen, J. Brandt *et al.*, "Transformer models in biomedicine," *BMC Medical Informatics and Decision Making*, vol. 24, p. 214, 2024.

[13] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Technical Report*, 2019.

[14] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, Z. Hu *et al.*, "Qwen technical report," *arXiv preprint arXiv:2309.16609*, 2023. [Online]. Available: https://arxiv.org/abs/2309.16609

[15] N. Gruver, M. Finzi, S. Qiu, and A. G. Wilson, "Large language models are zero-shot time series forecasters," 2024.

[16] M. Jin, S. Wang, L. Ma, Z. Chu, J. Y. Zhang, X. Shi, P.-Y. Chen, Y. Liang, Y.-F. Li, S. Pan, and Q. Wen, "Time-llm: Time series forecasting by reprogramming large language models," 2024. [Online]. Available: https://arxiv.org/abs/2310.01728

[17] 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.

[18] M. Fraiwan, L. Fraiwan, B. Khassawneh, and A. Ibnian, "A dataset of lung sounds recorded from the chest wall using an electronic stethoscope," *Data in Brief*, vol. 35, p. 106913, 2021.

[19] Z. K. Abdul and A. K. Al-Talabani, "Mel frequency cepstral coefficient and its applications: A review," *IEEE Access*, vol. 10, pp. 122 136–122 158, 2022.

[20] Y. Huang, H. Hou, Y. Wang, Y. Zhang, and M. Fan, "A long sequence speech perceptual hashing authentication algorithm based on constant q transform and tensor decomposition," *IEEE Access*, vol. 8, pp. 34 140–34 152, 2020.