



OPEN ACCESS

EDITED BY

Ahsan H. Khandoker,
Khalifa University, United Arab Emirates

REVIEWED BY

Xihe Qiu,
Shanghai University of Engineering
Sciences, China
Yuttapong Jiraraksoyakun,
King Mongkut's University of Technology
Thonburi, Thailand

*CORRESPONDENCE

Liuying Li,
✉ arenally@sina.com
Xia Zhou,
✉ zhoux1823@163.com

[†]These authors have contributed equally to
this work and share first authorship

RECEIVED 10 July 2025

REVISED 29 October 2025

ACCEPTED 31 October 2025

PUBLISHED 26 November 2025

CITATION

Hua C, Liu Z, Li L, Zhou X and Xiang C (2025)
Snoring sound classification in patients with
cerebrovascular stenosis based on an
improved ConvNeXt model.
Front. Physiol. 16:1661258.
doi: 10.3389/fphys.2025.1661258

COPYRIGHT

© 2025 Hua, Liu, Li, Zhou and Xiang. This is an
open-access article distributed under the
terms of the [Creative Commons Attribution
License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or
reproduction in other forums is permitted,
provided the original author(s) and the
copyright owner(s) are credited and that the
original publication in this journal is cited, in
accordance with accepted academic practice.
No use, distribution or reproduction is
permitted which does not comply with
these terms.

Snoring sound classification in patients with cerebrovascular stenosis based on an improved ConvNeXt model

Caijian Hua^{1†}, Zhihui Liu^{1†}, Liuying Li^{2*}, Xia Zhou^{2*} and
Caorong Xiang¹

¹School of Computer Science and Engineering, Sichuan University of Science and Engineering, Yibin, China, ²Traditional Chinese Medicine Department, Zigong First People's Hospital, Zigong, China

Introduction: Snoring is a common symptom of Obstructive Sleep Apnea (OSA) and has also been associated with an elevated risk of cerebrovascular disease. However, existing snoring detection studies predominantly focus on individuals with Obstructive Sleep Apnea-Hypopnea Syndrome (OSAHS), with limited attention given to the specific acoustic characteristics of patients with concomitant cerebrovascular diseases. To address this gap, this paper proposes a snoring classification method integrating dynamic convolution and attention mechanisms, and explores the acoustic feature differences between patients with cerebrovascular stenosis and those without stenosis.

Methods: First, we collected nocturnal snoring sounds from 31 patients diagnosed with OSAHS, including 16 patients with cerebrovascular stenosis, and extracted four types of acoustic features: Mel spectrogram, Mel-frequency cepstral coefficients (MFCCs), Constant Q Transform (CQT) spectrogram, and Chroma Energy Normalized Statistics (CENS). Then, based on the ConvNeXt backbone, we enhanced the network by incorporating the Alterable Kernel Convolution (AKConv) module, the Convolutional Block Attention Module (CBAM), and the Conv2Former module. We conducted experiments on snoring *versus* non-snoring classification and stenotic *versus* non-stenotic snoring classification, and validated the role of each module through ablation studies. Finally, the Mann-Whitney U test was applied to compare intergroup differences in low-frequency energy ratio, snoring frequency, and snoring event duration.

Results: This method achieves the best performance on the Mel spectrogram, with a snoring classification accuracy of 90.24%, compared to 88.16% for the ConvNeXt baseline model. It also maintains superiority in classifying stenotic *versus* non-stenotic snoring. Ablation analysis indicates that all three modules contribute to performance improvements. Moreover, the Mann-Whitney U test revealed significant differences ($p < 0.05$) between the stenotic and non-stenotic groups in terms of low-frequency energy ratio and nocturnal snoring frequency, whereas snoring event duration showed no significant difference.

Discussion: The proposed method demonstrates excellent performance in snoring classification and provides preliminary evidence for exploring acoustic features associated with cerebrovascular stenosis.

KEYWORDS

snoring sound classification, cerebrovascular stenosis, acoustic features, ConvNeXt, dynamic convolution, attention mechanisms

1 Introduction

1.1 Research background

Snoring is a breathing sound generated by the vibration of upper airway tissues that partially collapse during inhalation. It is common, occurring in approximately 50% of adults and 3.2%–12.1% of children (Ersu et al., 2004; Duckitt et al., 2006), and is widely regarded as an early clinical sign of obstructive sleep apnea-hypopnea syndrome (OSAHS) (Jiang et al., 2020), which causes damage to multiple systems throughout the body through intermittent hypoxia and sleep fragmentation. Research has established strong associations between OSAHS and various health conditions, including cardiovascular and cerebrovascular diseases (Zhu et al., 2024; Hong et al., 2022; Raptis et al., 2021), metabolic syndrome (Kargar et al., 2021), and cognitive impairment (Leng et al., 2016; Ylä-Herttuala et al., 2021). Notably, individuals with moderate to severe OSAHS face a stroke risk up to four times higher than that of the general population (Javaheri et al., 2022; Sanchez et al., 2022).

Studies have shown that the prevalence of OSAHS is significantly higher among patients with cerebrovascular diseases, such as stroke, than in the general population (Bassetti et al., 2006). Such patients may simultaneously exhibit upper airway obstruction and central respiratory regulation abnormalities (Tanayapong and Kuna, 2021), and their snoring sounds may present unique acoustic features. Clinical studies have demonstrated a positive correlation between snoring energy in the 652–1,500 Hz frequency band and the common carotid artery intima-media thickness (CCA-IMT) (Lee et al., 2016). Early epidemiological research also suggests a significant association between heavy snoring and carotid atherosclerosis (Lee et al., 2008). Therefore, exploring the acoustic differences between snoring sounds in patients with concomitant cerebrovascular stenosis and those without stenosis may provide new clues about the underlying pathological mechanisms. This highlights the importance of further integrating artificial intelligence (AI) methods into such research.

1.2 Related work

In recent years, deep learning (DL) has been extensively applied in snoring detection and recognition, primarily along two directions: multimodal fusion and lightweight portable detection. Regarding multimodal approaches, Li et al. (2025) employed the end-to-end audio classification framework DFNet to fuse patient medical history with physiological information using a multi-branch convolutional architecture, achieving an accuracy of 84.1% in a four-class OSAHS classification task. Qiu et al. (2026) proposed the Multimodal Integration and Missing Audio Reconstruction (MIMAR-OSA) model, which uses multimodal data fusion and missing modality reconstruction strategies to maintain high diagnostic accuracy even when certain signals are unavailable. Gu et al. (Bing, 2025) proposed a hierarchical Transformer model that integrates electroencephalogram (EEG) signals with semantic features of English listening comprehension, significantly enhancing the screening performance for obstructive sleep apnea (OSA) in noisy environments. Hu et al. (2025) proposed the information

bottleneck-based parallel CNN-Transformer network (IPCT-Net), which achieves parallel fusion of local and global features, outperforming traditional methods on home sleep test data.

In the realm of lightweight and portable solutions, Zhang R. et al. (2024) developed a lightweight model based on long short-term memory spiking neural networks (LSTM-SNN), optimized through threshold coding and an alternative gradient method, reaching an accuracy of 93.4%. The MinSnore model proposed by Dinh et al. (2025) combines an advanced lightweight network architecture with self-supervised learning pre-training to deliver outstanding performance in practical applications. Sillaparaya et al. (2025) proposed a transfer learning framework that combines MobileNetV3-Large with a modified SENet module, achieving favorable performance on Mel-spectrogram and MFCC inputs under a 10-fold cross-validation, and showing potential for portable sleep apnea monitoring. Hong et al. (2025) achieved high sensitivity and specificity in home settings by combining smartphone recordings with a Vision Transformer model, providing a viable solution for wearable and portable sleep apnea detection.

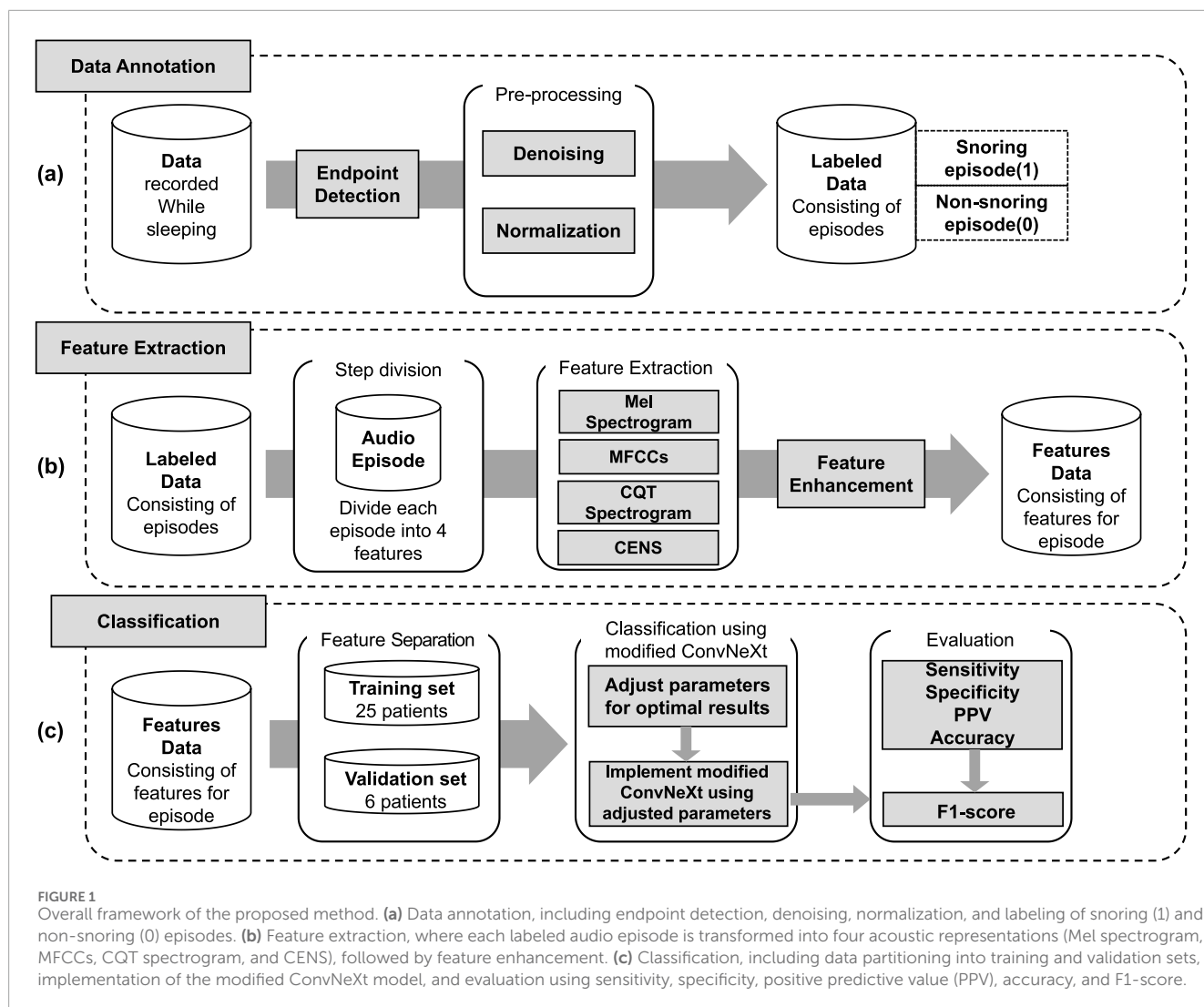
Although these approaches have contributed substantially to OSAHS diagnosis and snoring classification, two major limitations remain. First, most studies are conducted in controlled laboratory or home environments and lack validation in the complex acoustic environment of real hospital wards. Second, prior work has primarily focused on determining the presence or severity of OSAHS, and few have specifically analyzed patients with concomitant cerebrovascular stenosis as a distinct clinical subgroup.

In this study, we prospectively collected overnight snoring audio from patients diagnosed with OSAHS in a real hospital ward environment. Patients were grouped according to the presence or absence of concomitant cerebrovascular stenosis. The research objectives were as follows: (1) to construct a snoring and non-snoring classification model based on an improved ConvNeXt architecture, (2) to achieve classification between stenotic and non-stenotic snoring, and (3) to explore the acoustic feature differences between these two clinical subgroups and their clinical implications through statistical analysis.

2 Materials and methods

2.1 The overall framework of the proposed method

The proposed snoring classification method, incorporating dynamic convolution and attention mechanisms, consists of three main components: data annotation, feature extraction, and classification. As illustrated in Figure 1, snoring segments are labeled as 1, and non-snoring segments as 0. In the feature extraction phase, labeled audio segments are converted into four types of acoustic features: Mel spectrogram, Mel-frequency cepstral coefficients (MFCCs), Constant Q Transform (CQT) spectrogram, and Chroma Energy Normalized Statistics (CENS). After partitioning the extracted feature dataset into training and validation sets, the data was fed into an enhanced ConvNeXt model to perform the classification task of distinguishing snoring from non-snoring sounds. Furthermore, this framework was extended to classify snoring sounds between patients with cerebrovascular stenosis and



those without stenosis, exploring differences in snoring acoustic features across these distinct patient groups.

2.2 Data collection

This study collected snoring data from 31 patients at the Traditional Chinese Medicine Department, Zigong First People's Hospital, using a professional voice recorder and high-sensitivity microphone. Among them, 16 patients had cerebrovascular stenosis (6 mild, 5 moderate, 5 severe), while 15 patients had no stenosis. All patients were diagnosed with OSAHS through polysomnography (PSG). During hospitalization, each patient underwent continuous collection of 8-h nocturnal snoring data covering a full sleep cycle. To ensure data quality and diversity, microphones were positioned 2–3 cm from the patients' mouths, with signal collection conducted across various hospital environments. All patients underwent clinical evaluation and imaging examinations to confirm the presence or absence of cerebrovascular stenosis, and were subsequently categorized into "stenotic group" and "non-stenotic group" for analysis in follow-up studies.

2.3 Data annotation

All raw sleep sound signals were processed using endpoint detection to isolate valid acoustic segments, followed by noise reduction and normalization. Under the guidance of medical professionals, the segments were manually annotated in Audacity. The primary objective of this experiment was to identify snoring segments; thus, the signals were labeled into two categories: "snoring" and "non-snoring." Snoring segments were annotated based on a complete snoring cycle, ensuring each segment contained a full cycle. Non-snoring segments were selected from periods without snoring events, encompassing various sounds such as breathing, footsteps, coughing, and conversation. To reduce patient-dependent bias in model training and validation, the dataset was organized strictly at the patient level: 25 patients were assigned to the training set and 6 to the validation set. All segments from a given patient were included exclusively in one set, ensuring no segment-level mixing across training and validation. The overall training-to-validation ratio was maintained at 8:2, and this subject-independent strategy was applied consistently to both the "snoring *versus* non-snoring" classification and

TABLE 1 Dataset distribution.

Dataset	Snoring			Non-snoring			Overall total
	Stenotic	Non-stenotic	Total	Stenotic	Non-stenotic	Total	
Training (25 patients)	16,150	7900	24,050	15,850	9200	25,050	49,100
Validation (6 patients)	4018	1977	5995	4004	2336	6340	12,335
Total (31 patients)	20,168	9877	30,045	19,854	11,536	31,390	61,435

the “stenotic *versus* non-stenotic” snoring classification tasks. The distribution of audio segments across categories is shown in [Table 1](#).

2.4 Feature extraction

Previous studies ([Hou et al., 2024](#)) have demonstrated that audio spectrograms containing frequency and amplitude information that change over time can distinguish between different types of sounds. Building on this foundation, feature extraction was performed on manually annotated audio segments. Each clip was divided into 20-millisecond (ms) frames using a Hamming window with a 10-m frame shift, and all audio was uniformly resampled to 22.05 kHz using the Python library librosa. After zero-mean normalization, four types of feature maps were generated, with all images standardized to 224×224 pixels. Since the snoring event duration in this patient cohort generally lasted less than 3 s, audio was uniformly segmented into 3-s clips before conversion into images. Segments shorter than 3 s were padded with zeros to ensure consistent input length for the model.

2.4.1 Mel spectrogram

The Mel spectrogram extracts time-frequency features by simulating the auditory characteristics of the human ear. First, the audio is pre-emphasis, framing, and windowing. Pre-emphasis is the process of enhancing the high-frequency components of an audio signal using a high-pass filter. The signal is then segmented into short-time frames, and each frame is processed with a Hamming window to reduce spectral discontinuities at the boundaries. Next, the short-time Fourier transform (STFT) is computed for each frame to convert the time-domain signal into a frequency-domain representation ([Zhang R. et al., 2024](#)). The formula is shown in [Equation 1](#):

$$X(k) = \sum_{n=0}^{N-1} x(n) \cdot w(n) \cdot e^{-j2\pi kn/N}, \quad k = 0, 1, \dots, N-1 \quad (1)$$

In the formula: $x(n)$ represents the signal value of the n -th sampling point; $w(n)$ is the window function, which is the Hamming window here; N is the number of sampling points in each frame.

Finally, a Mel-scale triangular filter bank that converts linear frequencies to logarithmic frequencies is applied to the power spectrum to obtain the Mel spectrogram. This is achieved using the formula in [Equation 2](#):

$$s(m) = \ln \left(\sum_{k=0}^{N-1} |X(k)|^2 H_m(k) \right), \quad 0 \leq m \leq M \quad (2)$$

In the formula: $H(k)$ is the frequency response of the triangular filter, m denotes the m -th filter, and M is the number of Mel-scale triangular filter banks.

2.4.2 MFCCs

MFCCs further extract cepstral features based on the Mel spectrogram. After calculating the Mel spectrogram, the first 13 cepstral coefficients are obtained through discrete cosine transform (DCT) ([Zhang R. et al., 2024](#)), and the filter coefficients are decoupled. This process separates glottal excitation from vocal tract response, highlighting the static characteristics of the spectrum. The calculation formula is provided in [Equation 3](#):

$$C(l) = \sum_{m=0}^M s(m) \cos \left(\frac{\pi l(m-0.5)}{M} \right), \quad l = 1, 2, \dots, L \quad (3)$$

Where L represents the dimension of MFCCs, M denotes the number of Mel-scale triangular filter banks, identical to that in [Equation 2](#).

2.4.3 CQT spectrogram

CQT applies a logarithmic frequency scale to enable multi-resolution time-frequency analysis. The audio signal is first preprocessed through steps such as pre-emphasis and normalization. Subsequently, the CQT is computed to generate a time-frequency feature representation, where Q represents the Q factor, i.e., the ratio of the center frequency to the bandwidth ([Xie et al., 2021](#)). The Q factor is equal for all frequency intervals. The calculation of the Q factor is based on the two formulas in [Equations 4, 5](#):

$$Q = \frac{f_k}{\Delta f_k} \quad (4)$$

$$f_k = f_1 2^{\frac{k-1}{B}} \quad (5)$$

In the formulas: Δf_k is the bandwidth; f_k is the center frequency of the k -th bin; f_1 is the center frequency of the lowest frequency bin; B represents the number of bins within each octave frequency.

This method analyzes signals using a set of filters with logarithmic frequency distribution and constant Q values to generate complex spectrograms with high-frequency resolution.

2.4.4 CENS

CENS shows that chroma changes are more robust to temporal and timbre variations ([Abraham et al., 2023](#)). To extract CENS features, the audio signal is first processed through pre-emphasis, framing, windowing, and STFT. Then, the chroma information of

each frame is extracted through a 12-dimensional chroma filter bank. Finally, temporal smoothing and normalization are performed to generate the CENS feature representation.

2.5 Feature enhancement

The critical information in snoring sounds is often concentrated within specific frequency bands or short-term local regions. However, this information can be easily obscured by background noise or irrelevant components in the original feature maps, thereby reducing the model's ability to detect abnormal patterns. To address this issue, this paper draws on the enhanced spectrogram technique proposed by Wang et al. (2024) and proposes an adaptive filter bank enhancement method. Unlike methods that rely on fixed frequency band settings, the proposed method selects enhancement channels based on energy distribution, making it compatible with all four types of feature maps used in this study. Specifically, it evaluates the energy within each frame and dynamically adjusts filter response weights to emphasize high-energy regions. When the energy output of a filter exceeds the 80th percentile of that frame's energy distribution, an enhancement factor of β is applied; otherwise, the filter response remains unchanged. The β value was determined via grid search on the validation set of the baseline model in this paper ($\beta \in \{1.1, 1.2, \dots, 2.0\}$, step size 0.1), with $\beta = 1.5$ yielding the best performance on the validation set (see Supplementary Table S1). The enhanced filter bank is expressed in Equations 6–8:

$$H'_i(f, t) = a_i(t) \cdot H_i(f) \quad (6)$$

where,

$$a_i(t) = \begin{cases} \beta, E_i(P_t) > \text{percentile}_{80}(P_t) \\ 1, \text{otherwise} \end{cases} \quad (7)$$

and

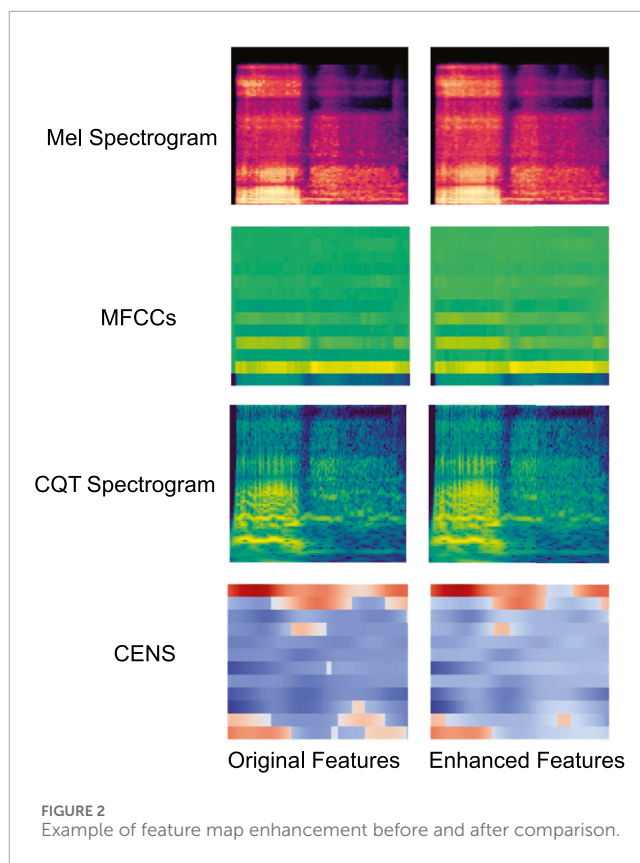
$$E_i(P_t) = \sum_f H_i(f) \cdot P_t(f) \quad (8)$$

In the formulas: $H_i(f)$ represents the response of filter i at frequency f ; $a_i(t)$ represents the enhancement factor of frame t ; β is the enhancement multiple; $P_t(f)$ represents the power spectrum of audio at frame t and frequency f ; $E_i(P_t)$ is the response value of filter i to P_t .

Figure 2 shows examples of visual comparisons of different feature maps before and after enhancement. The original feature maps are shown on the left, and the enhanced feature maps are shown on the right. The high-energy regions in the enhanced feature map are clearer and more prominent.

2.6 Classification

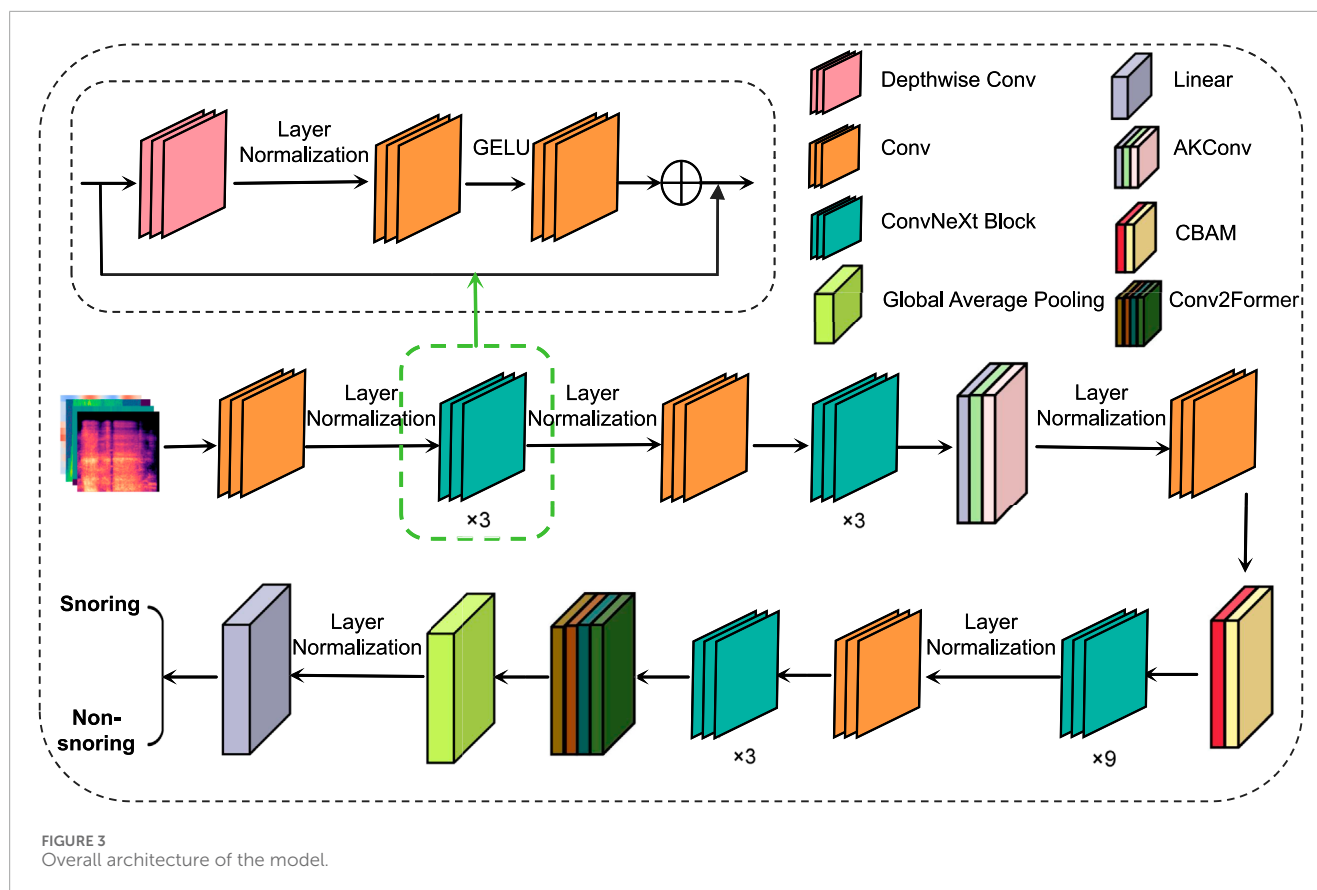
This paper proposes a snoring sound classification method that is based on an improved ConvNeXt model integrating dynamic convolution and attention mechanisms. The method consists of four components: the ConvNeXt network, the Alterable Kernel



Convolution (AKConv) module, the Convolutional Block Attention Module (CBAM), and the Conv2Former module. ConvNeXt serves as the backbone network, extracting multi-scale acoustic features. The AKConv module enhances adaptability to the irregular spectral features of snoring by dynamically adjusting the convolutional kernel shapes. CBAM utilizes a channel-spatial attention mechanism to emphasize key frequency bands while suppressing background noise. Finally, the Conv2Former module combines local and global feature interactions to effectively model the periodic dependencies of snoring events. This integrated design addresses snoring signals' characteristics of short-time non-stationarity, cross-band distribution, and periodicity. It enables the model to simultaneously capture both local spatiotemporal features and global periodic structures while enhancing robustness against noise interference, achieving superior feature representation capabilities compared to standalone modules or traditional ConvNeXt architectures. The overall model architecture is illustrated in Figure 3.

2.6.1 ConvNeXt

ConvNeXt (Liu et al., 2022), developed by Facebook AI Research in 2022, is a high-performance convolutional neural network that integrates key design ideas from Transformer architectures. By introducing large convolutional kernels, streamlining convolutional operations, and optimizing the overall network structure, ConvNeXt achieves substantial gains in both performance and computational efficiency. It is available in four variants: ConvNeXt-Tiny (T), Small (S), Base (B), and Large (L). In this study, the ConvNeXt-T model is adopted as the backbone network. The detailed structure is presented in Figure 4.



The input to the network is a feature map of size $224 \times 224 \times 3$. It first passes through an initial convolutional layer with a 4×4 convolutional kernel and a stride of four, generating a feature map of $56 \times 56 \times 96$, followed by layer normalization (LN) processing. This output then enters Stage 1, where it undergoes feature transformation through three ConvNeXt Blocks, each employing the standard structure of depthwise separable convolutions combined with multilayer perceptron (MLP). After downsampling with a stride of 2, a feature representation with a spatial size of 28×28 and 192 channels is obtained, which is then fed into Stage 2 comprising three ConvNeXt blocks. The Stage 2 output is downsampled to $14 \times 14 \times 384$ and input into Stage 3. After another downsampling step, Stage 3 generates a $7 \times 7 \times 768$ feature map, which is then fed into Stage 4. The resulting $7 \times 7 \times 768$ feature map is compressed by global average pooling into a $1 \times 1 \times 768$ vector, followed by LN. Finally, this vector is fed into a linear layer (also called fully connected layer) with an output dimension of 2 (where two represents the number of classification categories), yielding the classification results.

2.6.2 AKConv

In snoring sound classification tasks, traditional convolution operations struggle to effectively capture irregular snoring feature patterns in spectrograms. To address this, this study introduces the AKConv module (Zhang et al., 2023) into Stage 2 of ConvNeXt. Its core advantage lies in dynamically adjusting the shape and number of parameters of convolutional kernels to adaptively match the diverse morphological features of the snoring spectrum. In this study, the number of core parameters (num_param) was set to three

to balance model complexity and representational capacity, while the convolution stride was fixed at one to preserve spatial resolution. As illustrated in Figure 5, the AKConv module takes a spectrogram with dimensions (C, H, W) as input. It first determines the initial sampling position of the convolutional kernel through a coordinate generation algorithm. Specifically, the origin is fixed at the upper-left corner (0,0) as the common reference point for all kernel samples. The grid range is obtained by dynamically computing a base size (base_int) and handling the remainder: a regular grid region is first constructed, then the remaining points are added, resulting in an initial sampling layout that adapts to any number of parameters. Subsequently, a Conv2d layer is employed to learn spatial offsets of shape (B, 2N, H, W). The weights of this offset prediction layer are initialized to zero and undergo learning rate decay of 0.1 during backpropagation to ensure training stability. Subsequently, feature resampling is performed on the adjusted sampling points to precisely capture the deformed feature regions. Finally, the resampled features undergo dimension reshaping, convolution operations, and normalization processing. The optimized feature representations are then output via the SiLU activation function. It is worth noting that AKConv is an early variant of dynamic convolution. Its successor, LDConv (Zhang X. et al., 2024), further improves computational efficiency. However, AKConv was used in all experiments conducted in this paper.

2.6.3 CBAM

In this classification task, key frequency bands in the spectrogram (such as snoring harmonics) are often intermingled

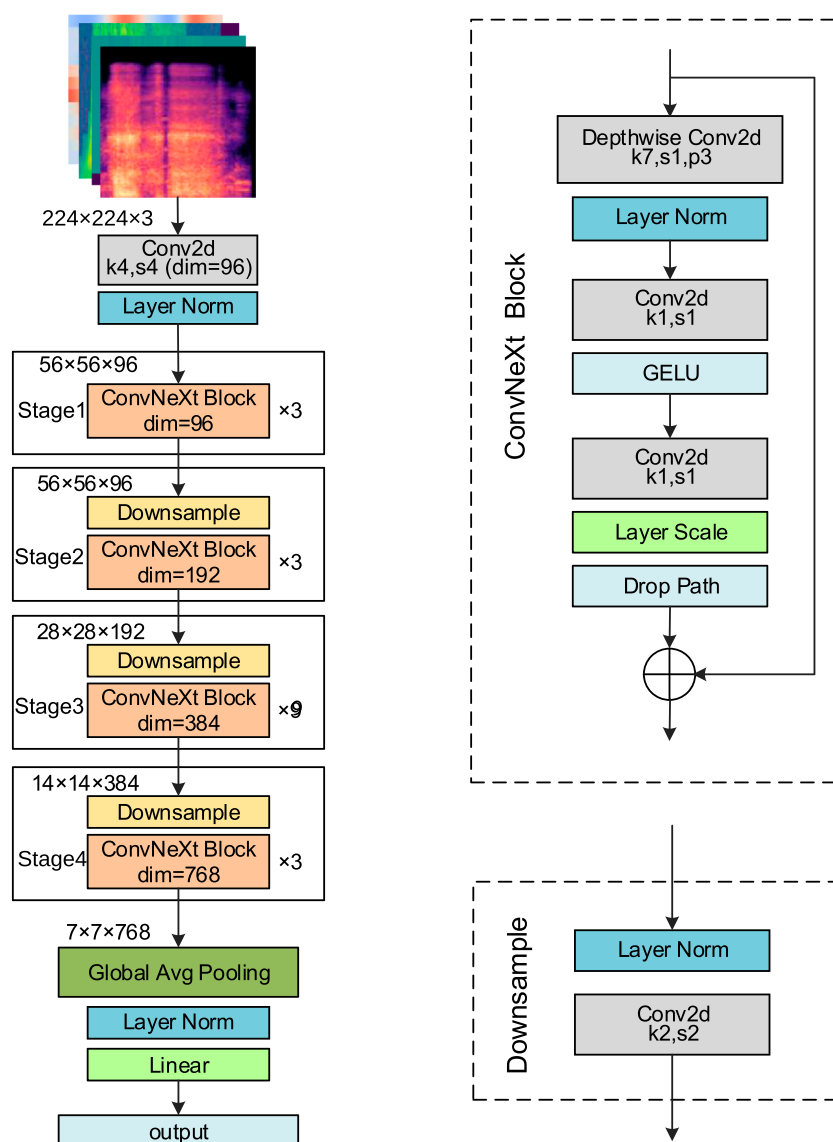


FIGURE 4
ConvNeXt-T network structure.

with background noise in both the spatial and channel dimensions. To address this specificity, this paper introduces the CBAM (Woo et al., 2018), which combines the Channel Attention Module (CAM) and the Spatial Attention Module (SAM). Its core function is to highlight important features through dual-path weighting, as shown in Figure 6.

The CAM performs max pooling and average pooling on the input feature map F (output size 1×1), respectively, passing through a shared two-layer MLP (with a reduction ratio of 16) to obtain feature maps A and B. The sum of these two feature maps is passed through a sigmoid function to generate the channel attention weight $M_C(F)$, which enhances key frequency bands associated with snoring sounds. The calculation formulas are given in Equations 9, 10:

$$M_C(F) = \sigma \left(W_1 \left(W_0 \left(F_{avg}^C \right) \right) + W_1 \left(W_0 \left(F_{max}^C \right) \right) \right) \quad (9)$$

$$F_1 = M_C(F) \odot F \quad (10)$$

In the formulas: $M_C(F)$ is the output weight of channel attention; σ is the sigmoid activation function; W_0 and W_1 are the weights of the first and second fully connected layers, respectively; F_{avg}^C and F_{max}^C represent average pooling and max pooling at the channel level, respectively; F_1 denotes the channel attention feature map obtained by weighting the original feature map F ; and $\odot$ denotes element-wise multiplication.

The SAM performs max pooling and average pooling on the feature map F_1 , then concatenates and fuses the results. It undergoes a 7×7 convolution with 3-stride padding to reduce the dimension to a single channel. Following this, a sigmoid function generates the spatial attention weights $M_S(F_1)$, focusing on regions with concentrated energy in the spectrum. The calculation formulas are presented in Equations 11, 12:

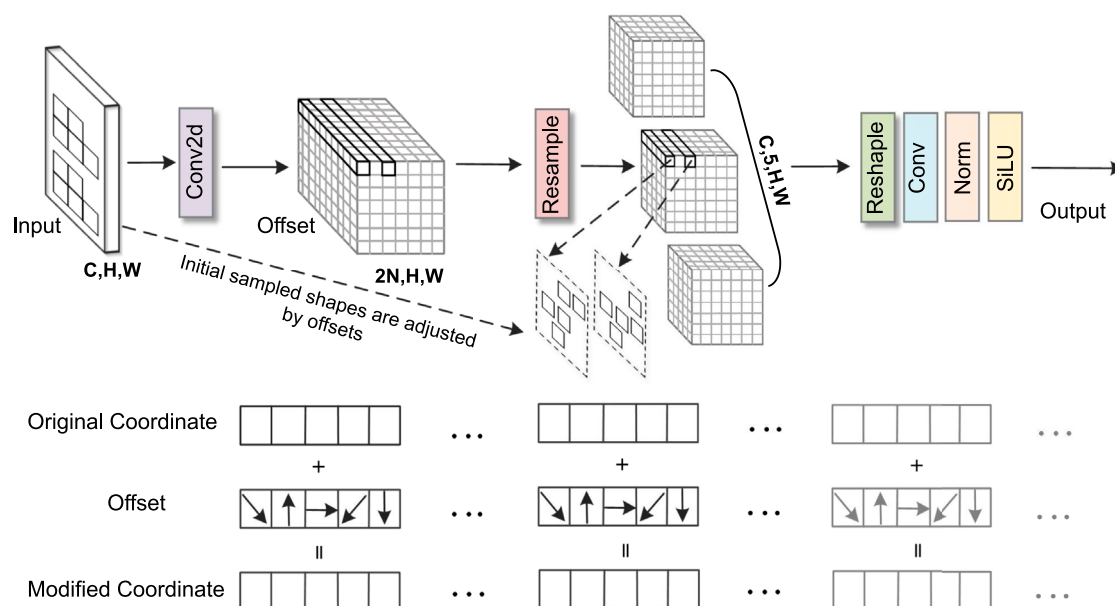


FIGURE 5
AKConv network structure.

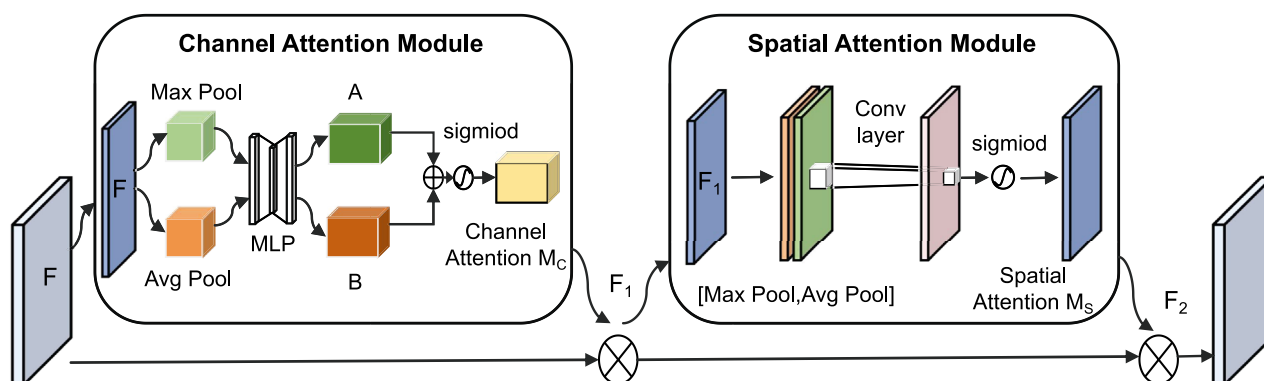


FIGURE 6
CBAM network structure.

$$M_S(F_1) = \sigma(f^{7 \times 7}([F_{avg}^S; F_{max}^S])) \quad (11)$$

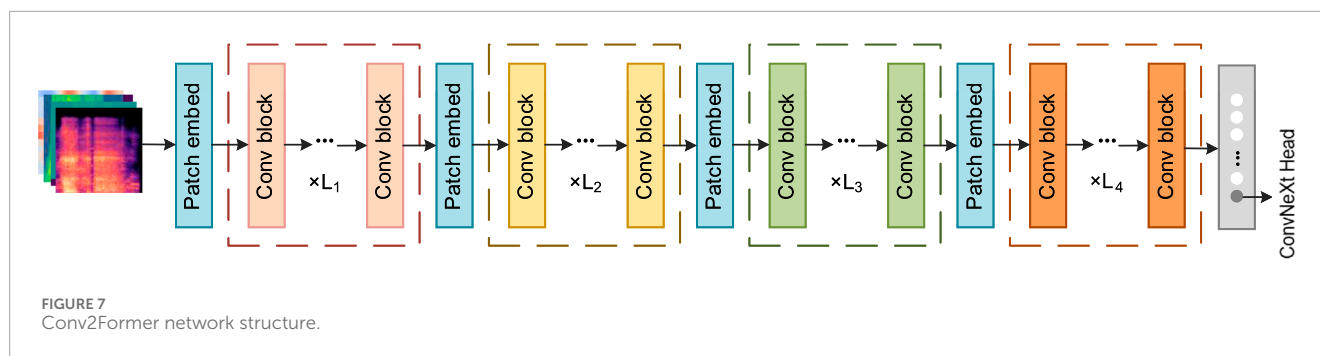
$$F_2 = M_S(F_1) \odot F_1 \quad (12)$$

In the formulas: $M_S(F_1)$ is the output weight of spatial attention; $f^{7 \times 7}$ is a 7×7 convolution kernel; F_{avg}^S and F_{max}^S are average pooling and max pooling at the spatial level, respectively; F_2 denotes the final output feature map of the CBAM module after spatial attention; and $\odot$ denotes element-wise multiplication.

2.6.4 Conv2Former

For the modeling requirements of long-range dependencies in snoring signals, traditional self-attention mechanisms are often employed. However, these methods typically incur high computational costs, which limit their practical application. To solve this problem,

this paper introduces the Conv2Former module (Hou et al., 2024) in Stage 4, whose structure is shown in Figure 7. This module is divided into four stages, each with different image resolutions, channel counts, and numbers of convolutional modulation blocks (Conv blocks). Adjacent stages reduce image resolution while increasing the number of channels through a patch embedding block (patch embed), employing a pyramid-like structure to extract features of different granularities. The Conv block replaces the self-attention layer in the Transformer with a convolutional modulation (ConvMod) layer. The key parameters of the ConvMod used in this study are as follows: the input and output dimensions are set to 768, matching the channel number of Stage 4; internally, an 11×11 depthwise separable convolution with padding of 5 is applied to preserve resolution, allowing the large convolutional kernel to encode spatial features and enhance global information modeling.



2.7 Statistical methods and 5-fold cross-validation

Statistical analyses were performed using SPSS version 26.0. Continuous variables were first tested for normality. Normally distributed data are expressed as mean $\pm$ standard deviation (SD) and compared between groups using the independent-samples t-test, whereas non-normally distributed data are expressed as median (interquartile range, IQR) and compared using the Mann-Whitney U test. Categorical variables were presented as counts and percentages, and group comparisons were conducted using the Chi-square (χ^2) test or Fisher's exact test when appropriate.

In ablation experiments, models were evaluated using 5-fold cross-validation based on patient grouping. Data from 31 patients were divided into five folds at the patient level, with each fold containing complete data from distinct patients (detailed statistics are provided in [Supplementary Table S2](#)). This ensured strict subject independence between folds and guaranteed that no segment-level information from the same patient appeared in both training and validation sets. In each iteration, one fold was used for validation and the remaining four for training. After repeating this process five times, each fold had been used as the validation set once. Results are presented as mean $\pm$ SD for each fold. Differences between the model in this study and models with added modules were compared using the Wilcoxon signed-rank test. The significance level was set at $p < 0.05$, with significant differences marked with “*” in the tables.

3 Results

3.1 Experiment setup

This experiment was conducted in the following software and hardware environment: the processor is an Intel® Xeon® Silver 4210R (2.40 GHz), equipped with an NVIDIA RTX 3080 (10 GB) graphics card, and the operating system is Ubuntu 20.04.6. The deep learning framework used is PyTorch 2.4.1 (CUDA 11.8), running in a Python 3.8.20 environment. During training, the batch size was set to 32, the AdamW optimizer was used for parameter updates, the initial learning rate was 2×10^{-4} , and the weight decay coefficient was 0.05. The model was trained for 100 epochs, using the cross-entropy loss function to supervise the optimization process of the binary classification task.

3.2 Evaluation criteria

This paper uses five metrics to evaluate model performance, including accuracy, sensitivity, specificity, positive predictive value (PPV), and F1-score. Accuracy represents the overall proportion of correctly classified samples and serves as a basic indicator of classification performance. Sensitivity measures the model's ability to correctly identify actual snoring events, while specificity reflects its capacity to correctly recognize non-snoring samples. PPV refers to the proportion of true snoring samples among those predicted as snoring. F1-score provides a comprehensive assessment of the model's robustness under imbalanced category conditions.

The calculation formulas are shown in [Equations 13–17](#):

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (13)$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (14)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (15)$$

$$\text{PPV} = \frac{TP}{TP + FP} \quad (16)$$

$$\text{F1-score} = \frac{2\text{PPV} \cdot \text{Sensitivity}}{\text{PPV} + \text{Sensitivity}} \quad (17)$$

Where TP , TN , FP , and FN represent true positive, true negative, false positive, and false negative, respectively.

3.3 Performance evaluation of feature enhancement

To verify the applicability and effectiveness of the feature enhancement method proposed in this paper on different feature maps, a comparison experiment was conducted between the Mel spectrogram and CQT spectrogram before and after enhancement using the method proposed in this paper. The comparison results are shown in [Table 2](#). The experimental results show that the two enhanced spectrograms are superior to the unenhanced ones in all evaluation metrics, indicating that the adaptive filter bank enhancement method proposed in this paper can effectively improve the representation ability of key information in the feature map.

TABLE 2 Comparison of results before and after feature enhancement (Unit: %).

Feature map	Accuracy	Sensitivity	Specificity	PPV	F1-score
Mel Spectrogram	89.10	88.55	89.73	89.05	89.08
Enhanced Mel Spectrogram	90.24	89.67	90.88	90.25	90.24
CQT Spectrogram	87.46	87.37	87.67	87.59	87.51
Enhanced CQT Spectrogram	88.35	88.32	88.72	88.35	88.37

PPV, positive predictive value.

3.4 Comparative experiment

3.4.1 Snoring versus non-snoring classification

To validate the effectiveness of the proposed method, we conducted comparative experiments between snoring and non-snoring sounds using the four enhanced acoustic feature maps against the following eight classical classification networks. The experimental results are shown in [Table 3](#).

For statistical robustness, classification accuracy was further reported with 95% confidence intervals (CIs) estimated by patient-level bootstrap ($B = 1,000$ resamples, seed = 42), where each patient was treated as the sampling unit to avoid segment-level leakage. As shown in [Table 3](#), the proposed method significantly outperformed the comparison models in all evaluation metrics for the four acoustic feature maps. Among them, the Mel spectrogram achieved the highest classification accuracy at 90.24% (95% CI: 85.31%–94.67%), followed by MFCCs, CQT spectrogram, and CENS. The overall performance of the lightweight models MobileNetV3 and DFNet was lower than that of medium-sized models such as ResNet50 and DenseNet121. The ConvNeXt baseline model outperformed the traditional CNN architectures DenseNet121 and ResNet50, with accuracy gains of 1.71% and 2.29%, respectively, on the Mel spectrogram. Compared to the ConvNeXt baseline model, our method achieved approximately 2% improvements in accuracy, sensitivity, specificity, PPV, and F1-score on the Mel spectrogram. Accuracy gains on MFCCs, CQT spectrogram, and CENS also ranged from 2% to 4%. Overall, our method maintained a consistent advantage across models of different scales, and its F1-score and accuracy remained closely aligned, supporting its robustness. We also evaluated performance at the patient level through majority voting across segments ([Supplementary Table S3](#)). In the validation set ($n = 6$ patients), the proposed model correctly classified all patients (100.00% accuracy), whereas the ConvNeXt baseline reached 83.33% (5/6). This indicates that the improvement observed at the segment level is preserved at the clinically relevant per-patient level.

[Figure 8](#) shows the comparison of model accuracy between the Mel spectrogram and CQT spectrogram on the validation set. With the exception of the Vision Transformer network, which exhibited performance instability, all other models achieved accuracy above 80%. This paper's method demonstrates optimal convergence stability on different feature maps. In summary, the effective fusion of dynamic convolution and attention mechanisms enhances the model's ability to discriminate the time-frequency features of snoring.

3.4.2 Stenotic versus non-stenotic snoring classification

To investigate the characteristic differences between snoring sounds from patients with cerebrovascular stenosis and those without stenosis, we selected one representative model from each scale (small, medium, and large) for classification experiments and comparisons. Similarly, accuracy was reported with 95% CIs estimated using patient-level bootstrap ($B = 1,000$ resamples, seed = 42) to ensure subject independence. As summarized in [Table 4](#), all metric values of the proposed model outperformed other baseline models, with an accuracy of 84.68% (95% CI: 76.03%–89.62%) on the Mel spectrogram. These findings indicate that the proposed method demonstrates potential in distinguishing between stenotic and non-stenotic snoring sounds; however, these results should be regarded as preliminary due to the limited cohort size.

To further validate the robustness of these findings at a clinically relevant scale, we conducted patient-level analysis, as summarized in [Supplementary Table S4](#). When predictions were aggregated per patient in the validation set ($n = 6$), the proposed model achieved 83.33% accuracy (5/6 patients correctly classified) and identified all stenotic patients with 100.00% sensitivity (4/4), compared with 66.67% accuracy (4/6) for the ConvNeXt baseline. This suggests that the method may prioritize stenotic patients with high sensitivity, even with a small cohort.

3.5 Ablation experiment

To evaluate the impact of each proposed module on the performance of the ConvNeXt baseline, ablation experiments were conducted by incorporating different modules onto the four enhanced feature maps for snoring versus non-snoring classification. The results are summarized in [Table 5](#). Performance varies by module and feature map. On the Mel spectrogram, the AKConv module demonstrates the most significant performance, achieving an accuracy of 89.62%, which represents a 1.46 percentage point improvement over the baseline ConvNeXt model, while also increasing specificity to 90.05%. This suggests that dynamically adjusting the shape of convolutional kernels enhances adaptability to irregular snoring spectra. Similarly, the AKConv module performs best on the MFCCs, improving specificity by 1.9%. The CBAM module excels on the CQT spectrogram, increasing sensitivity by 2.04%. The Conv2Former module contributes most significantly to

TABLE 3 Comparison of classification performance for snoring and non-snoring sounds (Unit: %).

Feature map	Model	Accuracy (95% CI)	Sensitivity	Specificity	PPV	F1-score
Mel Spectrogram	ViT-B (Dosovitskiy et al., 2021)	76.43 (65.37–82.91)	77.36	78.21	78.52	78.03
	MobileNetV3-L (Sillaparaya et al., 2025)	82.51 (73.23–89.57)	81.88	83.02	82.24	82.37
	DFNet (Li et al., 2025)	84.32 (77.11–90.07)	83.57	84.80	84.15	84.16
	Swin-T (Liu et al., 2021)	84.26 (76.48–90.81)	83.31	85.17	84.39	84.60
	XCiT-T (El-Nouby et al., 2021)	83.09 (74.02–89.96)	83.13	83.76	83.21	83.20
	ResNet50 (He et al., 2016)	85.87 (79.36–91.79)	84.79	86.78	86.73	85.85
	DenseNet121 (Huang et al., 2017)	86.45 (80.47–91.63)	82.37	86.69	86.25	86.11
	ConvNeXt-T (Liu et al., 2022)	88.16 (82.97–92.54)	87.85	88.24	88.19	88.07
	This paper	90.24 (85.31–94.67)	89.67	90.88	90.25	90.24
MFCCs	ViT-B (Dosovitskiy et al., 2021)	69.45 (58.92–78.33)	68.76	69.88	70.12	69.50
	MobileNetV3-L (Sillaparaya et al., 2025)	79.00 (70.12–86.41)	78.31	79.52	79.03	78.87
	DFNet (Li et al., 2025)	81.31 (74.05–87.29)	80.85	81.66	81.25	81.30
	Swin-T (Liu et al., 2021)	82.31 (75.19–88.77)	82.77	81.73	82.27	82.34
	XCiT-T (El-Nouby et al., 2021)	79.58 (71.47–86.97)	79.67	78.81	79.79	79.82
	ResNet50 (He et al., 2016)	83.57 (77.41–89.21)	83.50	83.32	83.61	83.62
	DenseNet121 (Huang et al., 2017)	84.15 (78.33–89.62)	84.57	84.06	84.25	84.27
	ConvNeXt-T (Liu et al., 2022)	86.84 (81.40–91.27)	87.49	86.32	86.89	86.89
	This paper	89.11 (83.77–93.52)	89.44	88.67	89.12	89.21
CQT Spectrogram	ViT-B (Dosovitskiy et al., 2021)	77.38 (67.90–85.73)	76.82	78.43	78.25	77.64
	MobileNetV3-L (Sillaparaya et al., 2025)	80.55 (72.30–87.92)	79.68	81.23	80.46	79.94
	DFNet (Li et al., 2025)	82.03 (75.12–87.86)	81.55	82.67	82.14	82.00
	Swin-T (Liu et al., 2021)	83.05 (76.05–89.40)	82.14	84.11	83.30	83.22
	XCiT-T (El-Nouby et al., 2021)	82.33 (74.50–88.98)	82.42	82.69	82.37	82.38
	ResNet50 (He et al., 2016)	84.05 (78.10–89.65)	83.97	84.81	84.52	84.12
	DenseNet121 (Huang et al., 2017)	84.52 (79.12–89.48)	84.70	84.38	84.52	84.53
	ConvNeXt-T (Liu et al., 2022)	86.14 (81.02–90.76)	85.61	86.65	86.13	86.15
	This paper	88.35 (83.77–92.80)	88.32	88.72	88.35	88.37
CENS	ViT-B (Dosovitskiy et al., 2021)	66.89 (55.97–76.52)	74.63	59.14	64.57	69.23
	MobileNetV3-L (Sillaparaya et al., 2025)	73.59 (63.84–81.97)	74.00	72.80	73.12	73.25
	DFNet (Li et al., 2025)	76.16 (68.10–83.94)	76.30	75.52	76.07	76.13
	Swin-T (Liu et al., 2021)	77.20 (69.02–83.98)	79.43	75.11	77.24	77.25

(Continued on the following page)

TABLE 3 (Continued) Comparison of classification performance for snoring and non-snoring sounds (Unit: %).

Feature map	Model	Accuracy (95% CI)	Sensitivity	Specificity	PPV	F1-score
	XCiT-T (El-Nouby et al., 2021)	77.15 (69.05–83.84)	76.67	76.70	76.71	76.68
	ResNet50 (He et al., 2016)	82.87 (77.10–88.24)	82.24	83.55	83.78	82.91
	DenseNet121 (Huang et al., 2017)	81.66 (75.98–87.32)	80.62	82.45	82.59	81.63
	ConvNeXt-T (Liu et al., 2022)	84.37 (79.01–89.40)	84.19	84.58	84.44	84.40
	This paper	87.94 (82.83–92.45)	87.05	88.94	87.88	87.96

The MobileNetV3-L results in this table refer only to the backbone, without the modified SENet, module used in (Sillaparaya et al., 2025). In this study, only the backbone was reproduced to ensure consistency in the comparative experiments. CI: confidence interval.

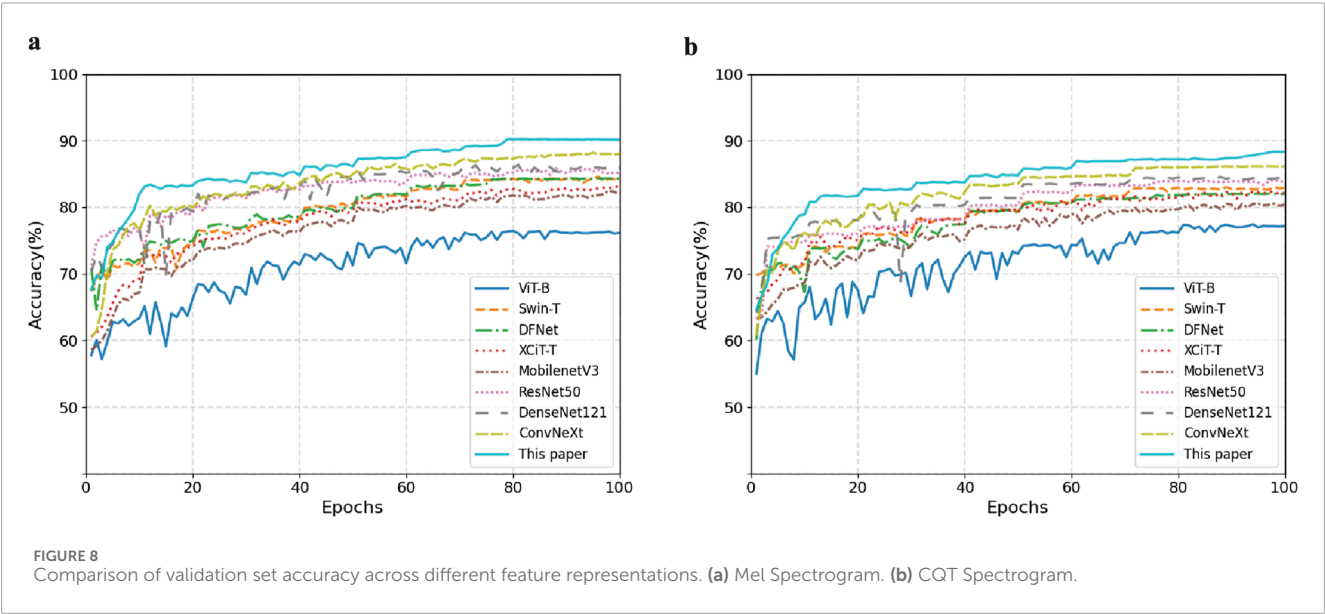


TABLE 4 Comparison of classification performance for stenotic and non-stenotic snoring sounds (Unit: %).

Feature map	Model	Accuracy (95% CI)	Sensitivity	Specificity	PPV	F1-score
Mel Spectrogram	ViT-B	69.58 (57.12–79.24)	69.30	69.92	69.55	69.35
	MobileNetV3-L	77.34 (66.20–85.92)	77.65	77.05	77.21	77.43
	ResNet50	81.92 (73.05–88.48)	81.40	82.18	81.85	81.70
	ConvNeXt-T	82.41 (74.22–88.95)	81.76	83.05	82.49	82.28
	This paper	84.68 (76.03–89.62)	84.12	84.35	84.92	84.97

the CENS map, improving accuracy by 1.97% and demonstrating its strength in global feature modeling.

To further validate the effectiveness of the proposed method, we plotted the loss curves of the baseline model and our model on both the training and validation sets. As shown in Figure 9, our method converges faster during training and achieves a lower final loss value, indicating that the introduced module enhances feature extraction capabilities and improves the model’s classification performance.

We further validated module performance using 5-fold cross-validation on the Mel spectrogram and assessed significance with the Wilcoxon test (Table 6). The ConvNeXt-T model achieved an average accuracy of 88.24%. Introducing different modules on this foundation yielded performance improvements, with AKConv demonstrating the most prominent enhancement. CBAM and Conv2Former also delivered gains, but the differences compared to our proposed model were statistically significant ($p < 0.05$).

TABLE 5 Ablation experiment results (Unit: %).

Feature map	Model	Accuracy	Sensitivity	Specificity	PPV	F1-score
Mel Spectrogram	ConvNeXt-T	88.16	87.85	88.24	88.19	88.07
	AKConv	89.62	89.10	90.05	89.58	89.62
	CBAM	89.11	88.65	89.45	89.06	89.08
	Conv2Former	88.75	88.40	88.98	88.80	88.77
	This paper	90.24	89.67	90.88	90.25	90.24
MFCCs	ConvNeXt-T	86.84	87.49	86.32	86.89	86.89
	AKConv	88.45	88.83	88.22	88.41	88.46
	CBAM	88.14	88.40	87.85	88.05	88.12
	Conv2Former	87.95	88.12	87.72	87.90	87.96
	This paper	89.11	89.44	88.67	89.12	89.21
CQT Spectrogram	ConvNeXt-T	86.14	85.61	86.65	86.13	86.15
	AKConv	87.10	86.95	87.24	87.06	87.00
	CBAM	87.92	87.65	88.20	87.90	87.88
	Conv2Former	87.35	87.40	87.28	87.34	87.33
	This paper	88.35	88.32	88.72	88.35	88.37
CENS	ConvNeXt-T	84.37	84.19	84.58	84.44	84.40
	AKConv	85.18	84.68	85.52	85.08	85.14
	CBAM	85.75	85.23	86.25	85.74	85.72
	Conv2Former	86.34	86.02	86.23	86.33	86.34
	This paper	87.94	87.05	88.94	87.88	87.96

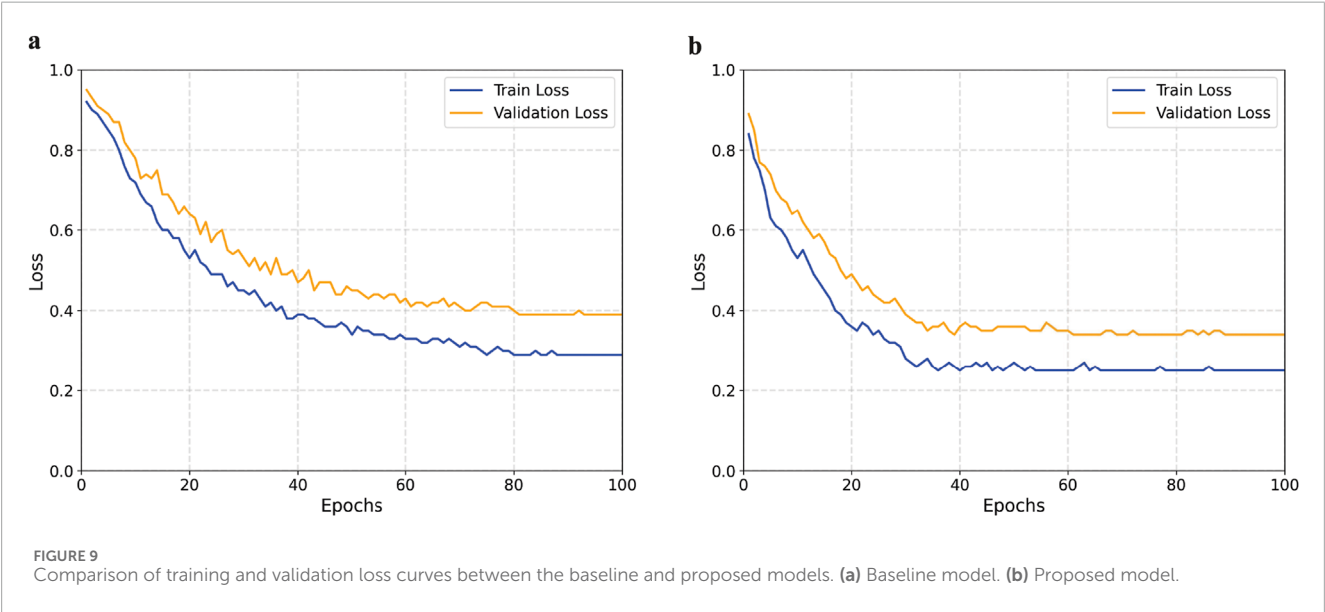


TABLE 6 Comparison of 5-fold cross-validation performance for ablation experiments on the Mel spectrogram (Unit: %).

Model variant	Accuracy	Sensitivity	Specificity	PPV	F1-score
ConvNeXt-T	88.24 ± 0.88*	87.91 ± 0.95*	88.32 ± 0.90*	88.27 ± 0.93*	88.19 ± 0.91*
+AKConv	89.57 ± 0.93*	89.12 ± 1.01*	90.03 ± 0.88*	89.51 ± 1.06*	89.59 ± 0.92*
+CBAM	89.18 ± 0.95*	88.69 ± 1.03*	89.56 ± 0.91*	88.87 ± 1.08*	89.11 ± 0.94*
+Conv2Former	88.76 ± 0.97*	88.33 ± 1.05*	88.01 ± 0.93*	88.42 ± 1.11*	88.78 ± 0.96*
This paper	90.31 ± 0.82	89.74 ± 0.95	90.88 ± 0.91	90.28 ± 0.97	90.34 ± 0.89

Results are reported as mean ± SD over 5-fold cross-validation. Values marked with * indicate significant differences compared with the proposed model, based on the Wilcoxon signed-rank test ($p < 0.05$).

TABLE 7 Computational efficiency comparison of different models.

Model	Params (M)	FLOPs (G)	GPU Latency (ms)	GPU Throughput (1mg/s)	CPU Latency (ms)	CPU Throughput (1mg/s)
ResNet50	23.51	4.13	12.05	82.98	4043	0.25
MobileNetV3	4.20	0.23	13.08	76.40	2900	0.34
Swin-T	27.52	3.13	39.41	25.37	2139	0.47
ConvNeXt-T	27.82	4.49	15.03	65.20	2200	0.45
This paper	32.05	5.20	16.12	61.80	2398	0.42

Params: number of parameters. FLOPs: floating-point operations per second. img/s: images processed per second.

This indicates that while each module contributes to feature enhancement, our model achieves the best overall performance.

3.6 Computational efficiency comparison analysis

To evaluate the computational efficiency, several representative models were compared on the dataset in this paper, as shown in Table 7. It can be observed that ResNet50 achieves the lowest latency of 12.05 m and the highest throughput of 82.98 images per second (img/s) on the GPU, but its latency increases significantly during CPU inference. MobileNetV3, as a lightweight model with minimal parameters and floating-point operations per second (FLOPs), performs relatively fast on the GPU but underperforms compared to ConvNeXt-T on the CPU. Although Swin-T has a large number of parameters, it exhibits the highest GPU latency and lowest throughput. ConvNeXt-T demonstrates a good balance on both GPU and CPU, with relatively stable latency and throughput. In comparison, the proposed model in this paper increases both parameters and computational cost, resulting in slightly lower inference latency and throughput than ConvNeXt-T. Nevertheless, it still significantly outperforms Swin-T and other lightweight models. It achieves the highest recognition accuracy while maintaining good computational efficiency, offering the best overall performance.

3.7 Noise robustness experiment

To evaluate the generalization ability of the proposed method in real-world environments, noise robustness experiments were conducted. Background noise was selected from the publicly available MUSAN dataset (Snyder et al., 2015), which contains speech, music, and various environmental sounds, to simulate interference in hospital and daily settings. For the snoring *versus* non-snoring classification validation set, each audio segment was linearly mixed with MUSAN noise at different signal-to-noise ratios (SNR = +20, +10, 0, −5 dB), with noise energy scaled to achieve the target SNR. The mixed audio segments were preprocessed following the same procedures as in the model training phase and converted into Mel spectrograms for model input.

Table 8 summarizes the performance of our method and the ConvNeXt-T baseline model under different noise conditions. Under mild noise (+20 dB), both models exhibit only a slight decline compared to the noise-free condition. As the noise level increases, our method shows a significantly slower performance degradation, reflecting its stronger resistance to interference. At 0 dB SNR, our model maintains an F1-score of 83.36%, significantly outperforming the baseline model. Even under extreme noise conditions (−5 dB), it still achieves an F1-score of 78.62%, exceeding the baseline model by over 8 percentage points. These results conclusively demonstrate that the introduced module enhances the model’s feature extraction and discrimination capabilities in complex acoustic environments,

TABLE 8 Performance comparison of proposed model and ConvNeXt-T under different SNR conditions.

Model	SNR (dB)	Performance (%)				
		Accuracy	Sensitivity	Specificity	PPV	F1-score
This paper	∞	90.24	89.67	90.88	90.25	90.24
	+20	89.12	88.75	89.53	88.87	88.82
	+10	86.65	85.20	87.25	85.59	85.87
	0	83.58	82.11	83.79	83.26	83.36
	−5	78.32	77.05	78.16	77.97	78.62
ConvNeXt-T	∞	88.16	87.75	88.24	88.19	88.07
	+20	86.20	85.89	86.95	86.13	85.94
	+10	82.67	81.52	82.89	81.92	82.05
	0	77.58	76.33	78.05	77.43	77.14
	−5	70.96	69.74	71.35	70.30	70.58

SNR: signal-to-noise ratio.

thereby improving its robustness and practicality in real-world scenarios.

3.8 Visualization experiment

In order to assess the feature extraction capability of the proposed model, Grad-CAM (Selvaraju et al., 2017) was employed to visualize the feature maps from the final layer of the feature extraction network. Four types of acoustic representations from a randomly selected snoring segment were used as input. Then, ViT-B, Swin-T, ResNet50, DenseNet121, and ConvNeXt networks were chosen for feature map visualization experiments. The results are shown in Figure 10. Taking the Mel spectrogram and CQT spectrogram as examples, our method demonstrates greater focus on the key frequency bands of snoring compared to the other five networks, whose focus appears more dispersed. This confirms that our approach captures more discriminative acoustic features.

3.9 Comparative analysis of stenotic and non-stenotic patients

3.9.1 Demographic characteristics

Beyond algorithmic research, we also conducted an in-depth analysis of the clinical characteristics of the study subjects. Table 9 summarizes the baseline demographic characteristics and comorbidities of patients in the stenotic and non-stenotic groups. No significant differences were observed between the two groups in terms of age, gender distribution, hypertension, diabetes, atherosclerosis, dyslipidemia, or apnea-hypopnea index (AHI) ($p > 0.05$). Given that cerebral infarction and cerebral ischemia are common clinical consequences of cerebrovascular stenosis, only

descriptive statistics were performed for these two medical histories. Overall, the two groups were comparable in primary demographic and vascular risk factors, providing a reliable foundation for analyzing snoring-related characteristics.

3.9.2 Analysis of differences in snoring acoustic features

We used the Mann-Whitney U test to analyze differences between the stenotic group and the non-stenotic group in three aspects: low-frequency energy ratio (<650 Hz energy proportion), snoring frequency, and snoring event duration. Snoring event duration was calculated by statistically determining the start and end points of each independent snoring event and measuring its duration; the median duration of all events per patient was used as the representative value. The low-frequency energy ratio was calculated based on the power spectral density (PSD) of each snoring segment. Specifically, the ratio of energy in the <650 Hz band to total energy was used as the low-frequency energy ratio for that segment. The median value across all segments for each patient was then taken as the individual indicator. The 650 Hz threshold referenced the 652 Hz threshold proposed by Lee et al. (Lee et al., 2016) and was rounded for computational simplicity.

Results shown in Table 10 indicate that the “low-frequency energy ratio” was significantly lower in the stenotic group than in the non-stenotic group ($p = 0.025$). This suggests that snoring energy distribution in stenotic patients tends toward high-frequency components, consistent with previous findings showing a positive correlation between snoring energy in the 652–1,500 Hz band and CCA-IMT (Lee et al., 2016). Regarding snoring frequency, the stenotic group exhibited significantly more nocturnal snoring events per 8-h sleep period compared to the non-stenotic group ($p = 0.031$). This likely reflects more frequent upper airway obstruction or turbulence, aligning with Lee et al.’s (Lee et al.,

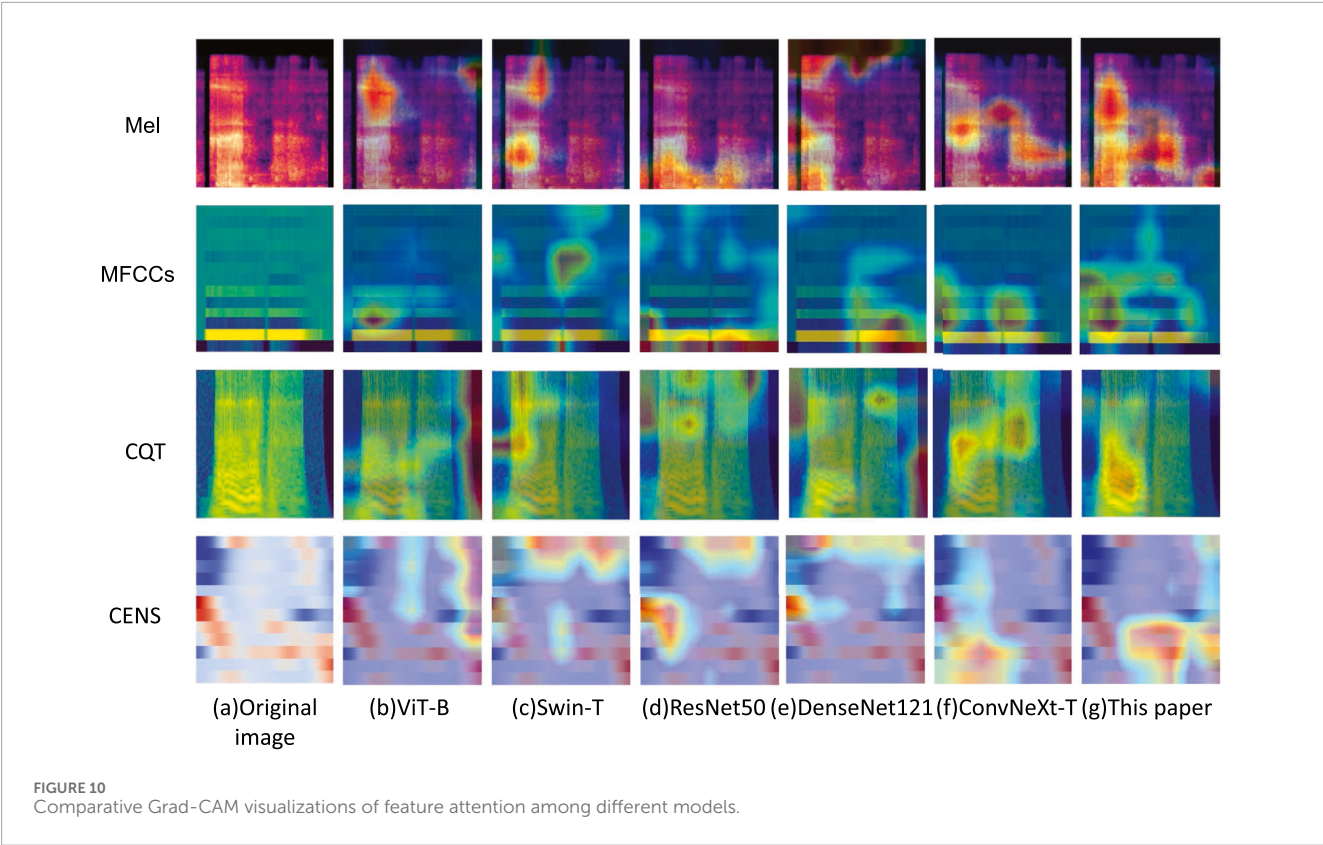


TABLE 9 Baseline characteristics of patients in stenotic and non-stenotic groups.

Item	Stenotic group (n = 16)	Non-stenotic group (n = 15)	t/χ^2	p-value ^a
Age (years)	71.44 ± 7.70	67.53 ± 9.12	1.284	0.210
Sex			0.819	0.366
Female	7 (43.8)	9 (60.0)		
Male	9 (56.2)	6 (40.0)		
Hypertension	11 (68.8)	9 (60.0)	0.259	0.611
Diabetes	8 (50.0)	6 (40.0)	0.313	0.576
Atherosclerosis	7 (43.8)	6 (40.0)	0.045	0.833
Dyslipidemia	6 (37.5)	5 (33.3)	0.000	1.000
AHI (events/h)	16.05 ± 5.90	13.75 ± 3.80	1.283	0.210
Cerebral infarction	4 (25.0)	0 (0.0)	—	—
Cerebral ischemia	3 (18.8)	0 (0.0)	—	—

Data are expressed as mean ± SD or number (%). ^aContinuous variables were compared using independent-samples *t*-test, and categorical variables were compared using the χ^2 test, as appropriate. AHI: apnea-hypopnea index.

2008) conclusion that “heavy snoring is significantly associated with carotid atherosclerosis.” In contrast, snoring event duration was slightly longer in the stenotic group but did not reach statistical significance ($p = 0.185$), likely due to the limited sample size

or substantial individual variability. In summary, while this study did not identify significant intergroup differences in snoring event duration, the variations in low-frequency energy ratio and snoring frequency provide preliminary acoustic evidence for assessing

TABLE 10 Comparison of snoring characteristics between stenotic and non-stenotic groups.

Group	Low-frequency energy ratio	Snoring frequency (events/8 h)	Snoring event duration (s)
Stenosis group (n = 16)	0.52 (0.46–0.59)	835.50 (667.50–1,341.50)	1.60 (1.48–1.82)
Non-Stenosis group (n = 15)	0.69 (0.58–0.77)	649.00 (434.00–764.50)	1.51 (1.37–1.65)
Z	–2.253	2.174	1.344
p-value	0.025	0.031	0.185

Data are expressed as median (IQR). p values are reported to three decimal places. Statistical comparisons were performed using the Mann-Whitney U test. Significance level: $p < 0.05$.

TABLE 11 Sensitivity analysis of snoring characteristics after excluding cerebral infarction/ischemia cases.

Group	Low-frequency energy ratio	Snoring frequency (events/8 h)	Snoring event duration (s)
Stenosis group (n = 12)	0.52 (0.46–0.57)	835.50 (697.75–1,341.50)	1.59 (1.48–1.82)
Non-Stenosis group (n = 15)	0.69 (0.58–0.77)	649.00 (434.00–764.50)	1.51 (1.37–1.65)
Z	–2.098	2.390	1.073
p-value	0.038	0.018	0.294

Data are expressed as median (IQR). p values are reported to three decimal places. Statistical comparisons were performed using the Mann-Whitney U test. Significance level: $p < 0.05$.

cerebrovascular stenosis risk through snoring acoustic features, offering potential clinical implications. Future research should incorporate larger samples and quantitative respiratory parameters from PSG for in-depth validation.

To control for potential confounding effects of a history of cerebral infarction or cerebral ischemia, a sensitivity analysis was conducted after excluding these patients. As shown in Table 11, the Mann-Whitney U test revealed that the low-frequency energy ratio in the stenotic group remained significantly lower than that in the non-stenotic group ($p = 0.038$), and the snoring frequency was significantly higher ($p = 0.018$); while the difference in snoring event duration between the two groups remained statistically insignificant ($p = 0.294$). These findings indicate that even after controlling for the confounding effects of cerebral complications, the characteristics of high-frequency shift in snoring energy distribution and increased snoring frequency in stenotic patients persist stably, further validating the reliability of the aforementioned analysis.

4 Discussion

The ConvNeXt model proposed in this study, which integrates dynamic convolution and attention mechanisms, outperformed baseline networks in snoring sound classification and maintained stable performance under noisy conditions. Statistical analysis revealed significant differences in specific acoustic features (low-frequency energy ratio and nocturnal snoring frequency) between the cerebrovascular stenosis and non-stenosis groups. These findings provide preliminary evidence that snoring acoustics may serve as potential indicators of cerebrovascular risk, although further validation in larger cohorts is required.

4.1 Clinical significance and application scenarios

Based on the current findings, the proposed method demonstrates potential clinical application value in three aspects. First, in community or outpatient screening settings, automated snoring analysis can serve as a low-cost, non-invasive auxiliary tool for the preliminary identification of suspected cases. For example, patients exhibiting a markedly elevated nocturnal snoring frequency (≥ 800 events during an 8-h sleep period) and a clearly reduced low-frequency energy ratio could be prioritized for carotid ultrasound or other vascular assessments. Second, the model maintained stable performance under noisy conditions, indicating its applicability in hospital wards and home sleep monitoring scenarios. This adaptability expands the potential for continuous and long-term observation of high-risk individuals outside laboratory environments. Third, the observed acoustic differences between the stenosis and non-stenosis groups suggest that snoring features may serve as early warning indicators of cerebrovascular abnormalities in patients with OSAHS and help clinicians identify which patients require further vascular evaluation.

Future studies should establish quantitative criteria for referral or further evaluation, such as defining what level of nocturnal snoring frequency or range of low-frequency energy ratios should be considered clinically significant. These criteria should be validated through larger, prospective cohort studies.

4.2 Limitations and future work

Despite the advantages demonstrated by the proposed method, several limitations remain. First, the sample size is limited to

only 31 patients, leading to insufficient representativeness that may compromise the generalizability of the findings. Future studies should expand into multicenter settings and include participants across different age groups. Second, the method relies on two-dimensional spectrograms and convolutional neural networks, which, although effective in capturing time-frequency features, entail relatively high computational costs. Previous studies have proposed snoring classification methods based on one-dimensional features, such as amplitude spectrum trend features (Sun et al., 2020) or representation learning based on auditory receptive fields (Hu et al., 2023). These approaches generally offer higher computational efficiency and deployment convenience but remain inadequate for modeling complex time-frequency patterns. Future research could explore integrating two-dimensional spectral analysis with one-dimensional feature extraction to balance discriminative power and computational efficiency. Finally, the acoustic differences identified in this study are based solely on cross-sectional statistical analysis, making it difficult to directly infer the underlying pathological mechanisms. Larger-scale longitudinal studies are needed to validate their clinical significance.

5 Conclusion

This paper proposes a snoring classification method integrating dynamic convolution with attention mechanisms, with a particular emphasis on exploring the acoustic differences between patients with and without cerebrovascular stenosis. The main conclusions and contributions are as follows: (1) Methodological improvements: Among four acoustic features and multiple baseline models, this paper proposes integrating AKConv, CBAM, and Conv2Former modules into the ConvNeXt backbone. On the Mel spectrogram, the method achieves 90.24% accuracy in classifying snoring *versus* non-snoring sounds, representing an improvement of approximately 2 percentage points over the ConvNeXt baseline. It also achieves 84.68% accuracy in classifying stenotic *versus* non-stenotic snoring sounds, which should be regarded as preliminary. (2) Robustness and ablation validation: Through noise robustness experiments, the proposed method was shown to maintain stable performance under varying noise conditions. Furthermore, ablation studies confirm that each added module contributes incrementally to the observed improvements, underscoring the reliability and interpretability of the proposed architecture. (3) Preliminary clinical insights: Clinical analyses identified significant differences between stenotic and non-stenotic patients in low-frequency energy ratio ($p = 0.025$) and nocturnal snoring frequency ($p = 0.031$). Sensitivity analyses excluding patients with cerebral infarction and cerebral ischemia yielded consistent results ($p = 0.038$ and $p = 0.018$, respectively). Snoring event duration did not differ significantly. These clinical-statistical findings are promising but remain preliminary and require validation in larger-scale, longitudinal studies before any clinical application.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by Zigong First People's Hospital. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

CH: Project administration, Methodology, Formal Analysis, Investigation, Writing – review and editing, Conceptualization, Funding acquisition, Supervision. ZL: Conceptualization, Writing – original draft, Data curation, Visualization, Methodology, Validation, Software, Investigation. LL: Resources, Data curation, Project administration, Funding acquisition, Writing – review and editing. XZ: Project administration, Data curation, Funding acquisition, Resources, Writing – review and editing. CX: Writing – original draft, Data curation.

Funding

The authors declare that financial support was received for the research and/or publication of this article. This research was supported by the National Natural Science Foundation of China (Grant No. 42471437), the Zigong Science and Technology Plan Project (Grant No. 2023ZC22), and the Zigong Key Science and Technology Plan (Collaborative Innovation Class of Zigong Academy of Medical Sciences) Key Project (Grant No. 2023YKYXT03).

Acknowledgements

The authors would like to thank the staff of the Traditional Chinese Medicine Department, Zigong First People's Hospital, for their assistance with patient recruitment and data annotation guidance. We also acknowledge the hospital's logistical support during data collection. Additionally, we are grateful to all patients who voluntarily participated in this study.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The authors declare that no Generative AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated

organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fphys.2025.1661258/full#supplementary-material>

References

- Abraham, J. T., Khan, A. N., and Shahina, A. (2023). A deep learning approach for robust speaker identification using chroma energy normalized statistics and Mel frequency cepstral coefficients. *Int. Journal of Speech Technology* 26, 579–587. doi:10.1007/s10772-021-09888-y
- Bassetti, C. L., Milanova, M., and Gugger, M. (2006). Sleep-disordered breathing and acute ischemic stroke: diagnosis, risk factors, treatment, evolution, and long-term clinical outcome. *Stroke* 37, 967–972. doi:10.1161/01.STR.0000208215.49243.c3
- Bing, F., and Fan, Y. (2025). Ai-enabled osa screening using eeg data analysis and English listening comprehension insights. *Frontiers Med.* 12, 1534623. doi:10.3389/fmed.2025.1534623
- Dinh, N. N., Bach, N. C., Bach, T. V., Chi, D. T. N., Cuong, D. D., Dat, N. T., et al. (2025). Implementing deep learning on edge devices for snoring detection and reduction. *Comput. Biology Med.* 184, 109458. doi:10.1016/j.compbiomed.2024.109458
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2021). An image is worth 16x16 words: transformers for image recognition at scale. Available online at: <https://arxiv.org/abs/2010.11929> (Accessed September 14, 2025).
- Duckitt, W., Tuomi, S., and Niesler, T. (2006). Automatic detection, segmentation and assessment of snoring from ambient acoustic data. *Physiol. Meas.* 27, 1047–1056. doi:10.1088/0967-3334/27/10/010
- El-Nouby, A., Touvron, H., Caron, M., Bojanowski, P., Douze, M., Joulin, A., et al. (2021). Xcit: cross-covariance image transformers. Available online at: <https://arxiv.org/abs/2106.09681> (Accessed September 14, 2025).
- Ersu, R., Arman, A. R., Save, D., Karadag, B., Karakoc, F., Berkem, M., et al. (2004). Prevalence of snoring and symptoms of sleep-disordered breathing in primary school children in istanbul. *Chest* 126, 19–24. doi:10.1378/chest.126.1.19
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 770–778. doi:10.1109/CVPR.2016.90
- Hong, Z., Ou, Q., Cheng, Y., Xu, Y., Fei, H., and Liu, H. (2022). Cardiac imageology changes in patients with mild obstructive sleep apnea without cardiovascular disease. *Sleep Breath.* 26, 595–604. doi:10.1007/s11325-021-02421-0
- Hong, J., Yang, S. K., Kim, S., Cho, S.-W., Oh, J., Cho, E. S., et al. (2025). Real-time snoring detection using deep learning: a home-based smartphone approach for sleep monitoring. *Nat. Sci. of Sleep* 17, 519–530. doi:10.2147/NSS.S514631
- Hou, Q., Lu, C.-Z., Cheng, M.-M., and Feng, J. (2024). Conv2former: a simple transformer-style convnet for visual recognition. *IEEE Trans. Pattern Analysis Mach. Intell.* 46, 8274–8283. doi:10.1109/TPAMI.2024.3401450
- Hu, X., Sun, J., Dong, J., and Zhang, X. (2023). Auditory receptive field net based automatic snore detection for wearable devices. *IEEE Journal of Biomed. Health Inf.* 27, 2255–2263. doi:10.1109/JBHI.2022.3164517
- Hu, S., Wang, Y., Liu, J., Cui, Z., Yang, C., Yao, Z., et al. (2025). Ipct-net: parallel information bottleneck modality fusion network for obstructive sleep apnea diagnosis. *Neural Netw.* 181, 106836. doi:10.1016/j.neunet.2024.106836
- Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Q. (2017). “Densely connected convolutional networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. doi:10.1109/CVPR.2017.243
- Javaheri, S., Peker, Y., Yaggi, H. K., and Bassetti, C. L. (2022). Obstructive sleep apnea and stroke: the mechanisms, the randomized trials, and the road ahead. *Sleep. Med. Rev.* 61, 101568. doi:10.1016/j.smrv.2021.101568
- Jiang, Y., Peng, J., and Zhang, X. (2020). Automatic snoring sounds detection from sleep sounds based on deep learning. *Phys. Eng. Sci. Med.* 43, 679–689. doi:10.1007/s13246-020-00876-1
- Kargar, B., Zamanian, Z., Hosseinabadi, M. B., Gharibi, V., Moradi, M. S., and Cousins, R. (2021). Understanding the role of oxidative stress in the incidence of metabolic syndrome and obstructive sleep apnea. *BMC Endocr. Disord.* 21, 77–78. doi:10.1186/s12902-021-00735-4
- Lee, S. A., Amis, T. C., Byth, K., Larcos, G., Kairaitis, K., Robinson, T. D., et al. (2008). Heavy snoring as a cause of carotid artery atherosclerosis. *Sleep* 31, 1207–1213. doi:10.5665/sleep/31.9.1207
- Lee, G.-S., Lee, L.-A., Wang, C.-Y., Chen, N.-H., Fang, T.-J., Huang, C.-G., et al. (2016). The frequency and energy of snoring sounds are associated with common carotid artery intima-media thickness in obstructive sleep apnea patients. *Sci. Rep.* 6, 30559. doi:10.1038/srep30559
- Leng, Y., McEvoy, C. T., Allen, I. E., and Yaffe, K. (2017). Association of sleep-disordered breathing with cognitive function and risk of cognitive impairment: a systematic review and meta-analysis. *JAMA Neurol.* 74, 1237–1245. doi:10.1001/jamaneurol.2017.2180
- Li, B., Qiu, X., Tan, X., Yang, L., Tao, J., Fang, Z., et al. (2025). An end-to-end audio classification framework with diverse features for obstructive sleep apnea-hypopnea syndrome diagnosis. *Appl. Intell.* 55, 427. doi:10.1007/s10489-025-06299-3
- Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021). “Swin transformer: hierarchical vision transformer using shifted windows,” in *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, 10012–10022. doi:10.1109/ICCV48922.2021.00986
- Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. (2022). “A convnet for the 2020s,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 11976–11986. doi:10.1109/CVPR52688.2022.01167
- Qiu, X., Wei, Y., Tan, X., Xu, W., Wang, H., Ma, J., et al. (2026). Mimar-osa: enhancing obstructive sleep apnea diagnosis through multimodal data integration and missing modality reconstruction. *Pattern Recognit.* 169, 111917. doi:10.1016/j.patcog.2025.111917
- Raptis, D. G., Sinani, O., Rapti, G. G., Papanikolaou, A., Dadouli, K., Ntellas, P., et al. (2021). Clinically silent small vessel disease of the brain in patients with obstructive sleep apnea hypopnea syndrome. *Diagnostics* 11, 1673. doi:10.3390/diagnostics11091673
- Sanchez, O., Adra, N., Chuprevich, S., and Attarian, H. (2022). Screening for osa in stroke patients: the role of a sleep educator. *Sleep. Med.* 100, 196–197. doi:10.1016/j.sleep.2022.08.024
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). “Grad-cam: visual explanations from deep networks via gradient-based localization,” in *Proceedings of the IEEE international conference on computer vision (ICCV)*. doi:10.1109/ICCV.2017.74
- Sillapara, A., Jiraksopakun, Y., Chamnongthai, K., and Bhatranand, A. (2025). Attention-based transfer learning for efficient obstructive sleep apnea (Osa) classification on snore sound. *IEEE Access* 13, 113879–113889. doi:10.1109/ACCESS.2025.3575203
- Snyder, D., Chen, G., and Povey, D. (2015). Musan: a music, speech, and noise corpus. Available online at: <https://arxiv.org/abs/1510.08484> (Accessed September 14, 2025).
- Sun, J., Hu, X., Chen, C., Peng, S., and Ma, Y. (2020). Amplitude spectrum trend-based feature for excitation location classification from snore sounds. *Physiol. Meas.* 41, 085006. doi:10.1088/1361-6579/abaa34
- Tanayapong, P., and Kuna, S. T. (2021). Sleep disordered breathing as a cause and consequence of stroke: a review of pathophysiological and clinical relationships. *Sleep. Med. Rev.* 59, 101499. doi:10.1016/j.smrv.2021.101499
- Wang, F., Yuan, X., Liu, Y., and Lam, C.-T. (2024). Lungnext: a novel lightweight network utilizing enhanced mel-spectrogram for lung sound classification. *Journal of King Saud University-Computer Inf. Sci.* 36, 102200. doi:10.1016/j.jksuci.2024.102200

- Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S. (2018). "Cbam: convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*. doi:10.1007/978-3-030-01234-2_1
- Xie, J., Aubert, X., Long, X., van Dijk, J., Arsenali, B., Fonseca, P., et al. (2021). Audio-based snore detection using deep neural networks. *Comput. Methods Programs Biomed.* 200, 105917. doi:10.1016/j.cmpb.2020.105917
- Ylä-Herttuala, S., Hakulinen, M., Poutiainen, P., Laitinen, T. M., Koivisto, A. M., Remes, A. M., et al. (2021). Severe obstructive sleep apnea and increased cortical amyloid- β deposition. *Journal of Alzheimer's Dis.* 79, 153–161. doi:10.3233/JAD-200736
- Zhang, X., Song, Y., Song, T., Yang, D., Ye, Y., Zhou, J., et al. (2023). Akconv: convolutional kernel with arbitrary sampled shapes and arbitrary number of parameters. Available online at: <https://arxiv.org/abs/2311.11587> (Accessed September 14, 2025).
- Zhang, R., Li, R., Liang, J., Yue, K., Li, W., and Li, Y. (2024a). Long short-term memory spiking neural networks for classification of snoring and non-snoring sound events. *Chin. Journal of Electron.* 33, 793–802. doi:10.23919/cje.2022.00.210
- Zhang, X., Song, Y., Song, T., Yang, D., Ye, Y., Zhou, J., et al. (2024b). Ldconv: Linear deformable convolution for improving convolutional neural networks. *Image Vis. Comput.* 149, 105190. doi:10.1016/j.imavis.2024.105190
- Zhu, Y., Zhuang, Z., Lv, J., Sun, D., Pei, P., Yang, L., et al. (2024). Causal association between snoring and stroke: a Mendelian randomization study in a Chinese population. *Lancet Regional Health–Western Pac.* 44, 101001. doi:10.1016/j.lanwpc.2023.101001