We sincerely appreciate the valuable comments and suggestions from Reviewer #1, which have significantly improved the quality of our manuscript. In response, we have conducted thorough analyses, provided detailed explanations, and made corresponding revisions. The point-by-point responses to all comments are provided below.

Reviewer #1:

This paper presents Infra-Net, a parallel dual-branch architecture that integrates log-scattering wavelet transforms with a confidence-based fusion mechanism for infrasound classification. The approach is methodologically sound and physically motivated, with the column-wise multi-view feature treatment demonstrating a notable degree of novelty. The results on the public dataset are impressive; however, the significant accuracy drop observed on real-world CTBTO data raises concerns regarding model generalization and overfitting. The manuscript would benefit from a more rigorous ablation study and a detailed analysis of computational complexity. Therefore, author should make major revisions before considering potential publication in this journal.
Specific comments are as below:

1. Please clarify the novelty of "multi-view" input, authors should explicitly contrast this with prior work that concatenates features, and justify why treating columns independently is theoretically superior rather than an arbitrary data expansion.

Authors: Thank you very much for your suggestion. We have further elaborated on the novelty of the multi-view input. The multi-view concept in this paper is reflected not only in the input itself but also in the multi-view feature extraction from the input. The detailed explanation is as follows:

Regarding the multi-view input, its novelty lies in the column-wise independent processing of wavelet scattering coefficients (scattering path perspective). In contrast to conventional approaches in wavelet scattering transform (WST)-based signal classification (e.g., Lone & Aydin 2023; Priya et al. 2021), which usually concatenate the extracted multiple columns of scattering coefficients along the feature dimension into a single high-dimensional vector and then feed it into a single classifier, such practice "mixes" information from all scattering paths and ignores the physical independence of each column. The novelty of our work is that we take the logarithm of each column of scattering coefficients (each corresponding to a specific wavelet scale and scattering path) and treat it as an independent feature view for input. This is because the coefficients produced by the

wavelet scattering network at each layer and scale have clear physical meanings: some columns reflect the low-frequency energy envelope of the signal, some capture medium-to-high frequency transient details, and others provide stability against translation or small deformations. Forcibly concatenating these physically distinct columns not only introduces inter-channel redundancy and interference but also loses critical discriminative information due to local aggregation in subsequent convolution or pooling operations (the results in Table 9 show that our method outperforms feature concatenation approaches). In contrast, keeping each column independent (1) preserves the physical interpretability of each scattering path; (2) avoids masking subtle class-specific differences by feature mixing (since the distinction between two event types may only appear in one or two specific paths, and the logarithmic transformation further amplifies feature differences-the effectiveness of this step is fully demonstrated by the results in Table 8); and (3) reduces the probability of information loss, thereby enhancing feature representation capability. Moreover, it provides multiple independent prediction probabilities for the subsequent confidence-based fusion module, allowing it to learn the reliability of each view and thus achieve more robust ensemble decisions.

Regarding the multi-view feature extraction, its novelty lies in the complementary perspective of the dual-branch heterogeneous network (model architecture perspective), i.e., learning features from the same data using two completely different views: spatial multi-scale features (MSCI-Net) and temporal dependencies (GA-BiGRU). Specifically, MSCI-Net uses multi-scale convolutional kernels to capture the spatial-frequency distribution patterns of infrasound signals. Small-scale kernels detect local fluctuations, while large-scale kernels capture energy envelope trends. Special attention is given to structural differences among different events in the time-frequency representation (as shown in Figure 5). GA-BiGRU, in contrast, models long-range temporal dependencies using a bidirectional gated recurrent unit combined with a global attention mechanism. This design mitigates information decay that occurs in later hidden states of traditional recurrent networks. It also focuses on extracting dynamic patterns, such as sustained energy release in earthquakes and rapid decay in explosions. The novelty of this approach lies in acknowledging that, given the non-stationary, long-correlated, and multi-scale nature of infrasound signals, a single-branch network (whether pure CNN or pure RNN) struggles to simultaneously capture both the temporal evolution and the frequency-domain energy distribution. Our dual-branch structure acts as two "view-specific" feature extractors, each focusing on different physical attributes of the

signal. The ablation study (Table 7) also demonstrates that the accuracy of any single branch (Ma or Mb) is lower than that of simple dual-branch fusion (Mc), and our confidence-based fusion (Infra-Net) further outperforms simple averaging, proving the necessity of dual-view complementarity.

Finally, the two branches each generate multiple sets of probabilities from all scattering paths. These are first locally fused by averaging across paths, and then the joint confidence is computed via an inner-product operation. The core idea of this design is that the final decision favors a category only when both views show high confidence in that category, which is more stringent than simple probability addition or averaging, and can effectively suppress misjudgments or noise interference from a single branch. In summary, the "multi-view learning" in this paper covers three levels: data view (column-wise independent processing), model view (heterogeneous dual-branch), and decision view (confidence-based fusion). It has clear physical motivation and theoretical advantages.

We thank the reviewer for this insightful comment again, which has made our exposition of the essence of multi-view learning more rigorous and complete.

2. Page 21, Line 539: Please provide statistical significance testing for the CTBTO results. The reported 82.07% accuracy on CTBTO data is only marginally higher than simpler architectures (e.g., Mc at 81.89%).

Authors: Thank you very much for your important comments on statistical significance. In response to the experimental results shown on Page 21, Line 539 (corresponding to Table 10 in the original manuscript) and on Page 17, Line 430 (corresponding to Table 7 in the original manuscript), we have supplemented the corresponding statistical significance tests to more rigorously evaluate the performance differences between our proposed method (Infra-Net) and existing methods as well as ablation models (especially Mc).

First, all experimental results in Table 10 were obtained on the CTBTO field-measured dataset based on a three-fold event-independent split (as shown in Table 4 in the original manuscript). For each of the three subsets, multiple training and testing runs were conducted, and the final results were averaged. This table contains two types of comparisons: (1) comparison between our proposed Infra-Net and several existing wavelet scattering transform (WST)-based methods (WST-LSTM,

WST-BiLSTM, WST-BiGRU); (2) comparison where the MSCI-Net branch (specifically designed by us for infrasound signal characteristics) in Infra-Net was replaced by classical deep networks such as Alex-Net, VGG-16, and VGG-19, to validate the effectiveness of our lightweight multi-scale convolutional design. To assess the statistical significance of the differences between each comparison model and Infra-Net, we performed paired t-tests on the *ACC*, *P*, and *F1* score of each three-fold experiment. The results show that compared with the three existing WST-based methods, Infra-Net achieves average improvements of 4.54%, 3.79%, and 4.36% in *ACC*, *P*, and *F1*, respectively. All paired comparisons yielded p-values less than 0.05, indicating that these improvements are statistically significant. This fully validates the superiority of our method in multi-view input representation (column-wise independent processing of scattering paths) and multi-view feature extraction. On the other hand, after replacing MSCI-Net with Alex-Net, VGG-16, and VGG-19, the average performance of the three models decreased by 1.71%, 2.28%, and 1.06% in *ACC*, *P*, and *F1*, respectively, compared with Infra-Net. The paired t-test p-values were also all less than 0.05, demonstrating that our lightweight multi-scale convolutional network structure, specifically designed for small-sample infrasound signals, significantly outperforms traditional deep networks and can learn robust features more effectively from limited data.

Second, regarding the ablation experiment results in Table 7, we focus on the comparison between Infra-Net and the simple averaging fusion (Mc). All results in Table 7 are also based on the average of the three-fold event-independent split on the CTBTO dataset. Numerically, the accuracy of Infra-Net is only 0.18 percentage points higher than that of Mc. Although the improvement is modest, it still reflects the effectiveness of the confidence-based decision-making module proposed in this paper. The design rationale of this module is to emphasize the "joint support" of the two branches for the same category through an inner-product operation, which theoretically can suppress the risk of misjudgment by a single branch. This is confirmed by the experimental results. More importantly, compared with the traditional baseline method, Infra-Net achieves significant improvements of 6.34%, 8.56%, and 4.85% in *ACC*, *P*, and *F1*, respectively, with paired t-test p-values all far below 0.05. This substantial improvement is directly attributable to the multi-view learning strategy proposed in this paper, fully demonstrating the effectiveness of the overall framework.

To facilitate intuitive understanding of the above statistical test results, we have added a

supplementary table (**Table R1**) in the revised manuscript, summarizing the significance test results between our method and the closest-performing models such as Mc and the model replacing MSCI-Net with VGG-16. The results show that the performance improvement of our method is statistically significant across multiple repeated experiments.

Once again, we thank the reviewer for your attention to statistical rigor, which has prompted us to more objectively evaluate the advantages and limitations of our proposed method and has pointed the way for future improvements.

Table R1 The significance test results between Infra-Net and the closest-performing models such as Mc and the model replacing MSCI-Net with VGG-16.

Model	Source	Performance Metric	p-value	Significant?
Infra-Net vs Mc	Table 7	*ACC/P/F1*	$p < 0.05$	Yes
Infra-Net vs Mb	Table 7	*ACC/P/F1*	$p < 0.05$	Yes
Infra-Net vs Ma	Table 7	*ACC/P/F1*	$p < 0.05$	Yes
Infra-Net vs Replacing MSCI-Net with VGG-16	Table 10	*ACC/P/F1*	$p < 0.05$	Yes

3. Authors should discuss how to address the substantial generalization gap between datasets, because the performance drop from 100% (LOTIS) to 82.07% (CTBTO) suggests potential overfitting to the LOTIS acoustic environment.

Authors: Thank you for raising this important question regarding the classification accuracy achieved on the two different datasets used in our study. Regarding the classification accuracy of 82.07% and 100% on the two datasets, we provide the following explanations.

The lower classification accuracy on the first dataset can be attributed to three main factors:

(1) Initial analysis of misclassified signals reveals their predominant origin from distant infrasound events. As illustrated in Fig. 1, our statistical analysis of source distances in the CTBTO infrasound dataset shows that the nearest event source exceeds 800 kilometers. This long-range propagation introduces substantial variations in transmission paths and background noise across events. Signals become increasingly susceptible to noise interference, even to the point of being completely submerged, which diminishes the discernibility of source characteristics. PMCC detection results for misclassified signals (Fig. 2) confirm their significantly attenuated feature information compared to near-field events (Fig. 3), attributable to excessive propagation distances. Consequently, the model struggles to effectively learn

discriminative event features during training, ultimately reducing classification precision.

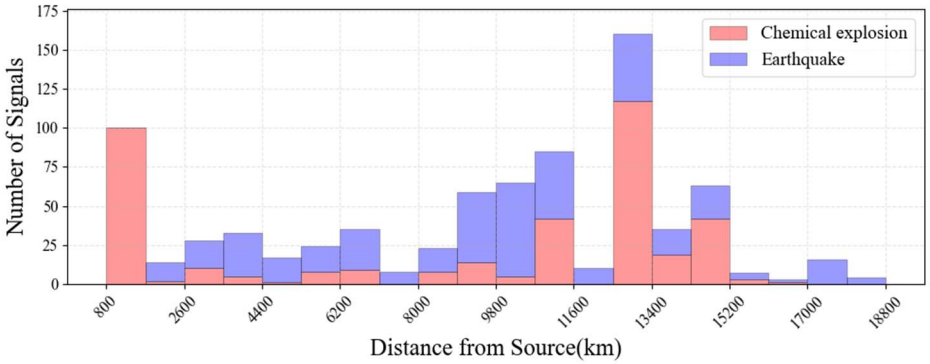

Fig. 1: The distribution of source distances for two types of infrasound event signals.

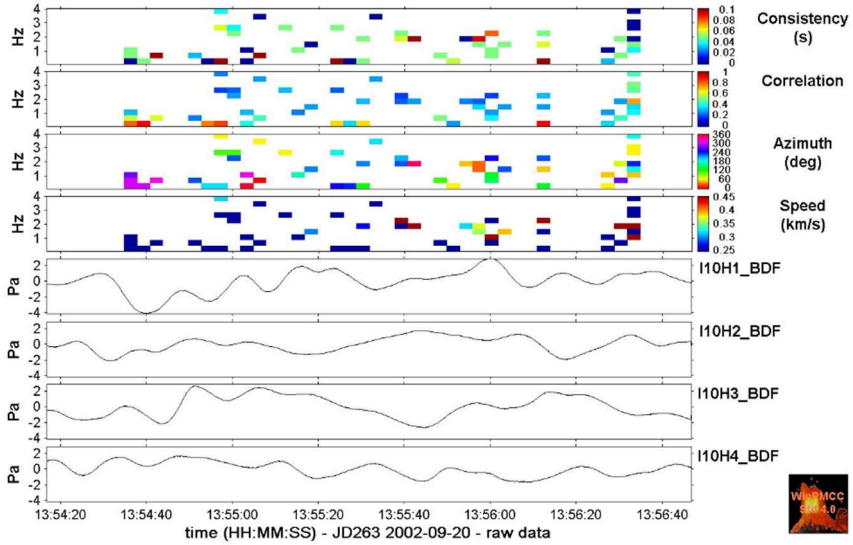

Fig. 2: The results of PMCC detection for misclassified infrasound signal.

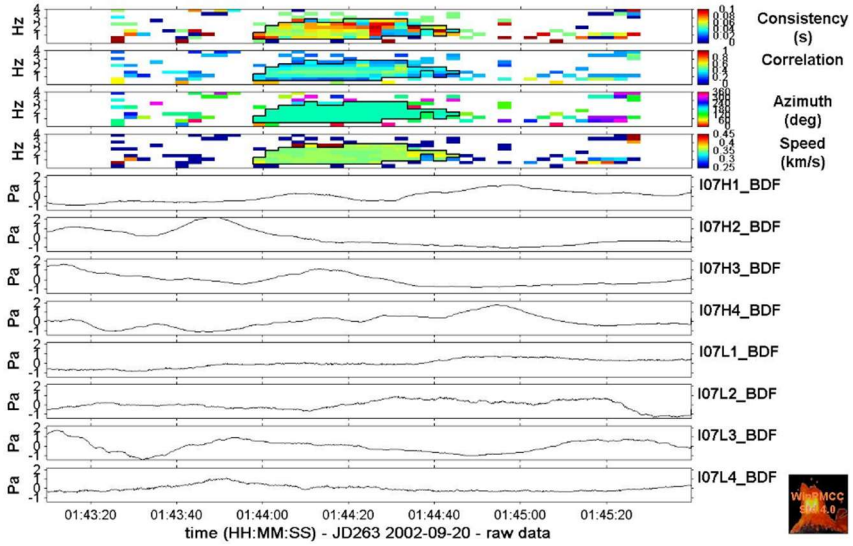

Fig. 3: The results of PMCC detection for near-field infrasound signal.

(2) The dataset includes chemical explosion and natural earthquake infrasound events. Earthquakes are continuous geological processes that generate infrasound signals with broad frequency ranges and long durations. Similarly, chemical explosions, especially continuous ones, produce infrasound signals with overlapping frequency ranges and durations. The similarity in signal generation mechanisms, along with attenuation and dispersion during propagation, causes overlap in frequency, amplitude, and waveform features of the two event types, making them difficult to distinguish effectively.

(3) To enhance model generalization, we strictly assigned signals from the same event to the same subset during dataset division. This further prevents overfitting due to learned event correlations. However, with the limited training sample size, this approach can result in even lower classification accuracy.

The second dataset achieved high classification accuracy. This is due to the differences from the first dataset:

(1) The distance from the event source to the monitoring station is small. The infrasound data mainly comes from the infrasound sensor array located in Windless Bight, Antarctica. The small source distance results in a high SNR of the received signals, which is beneficial for event classification. The other two datasets use the same monitoring equipment for signal recording, so they have good consistency.

(2) The differences in signal features of different types of events in the second dataset are more significant compared with the first dataset. For example, aurora-induced atmospheric gravity infrasound waves have unique frequency and propagation characteristics. Their signals usually have long-duration and low-frequency features. In contrast, infrasound waves generated by volcanic eruptions have high amplitude and long duration. These unique signal features, combined with the Infra-Net proposed in this paper, which has strong feature extraction and integrated decision-making capabilities, enable more accurate identification and classification of different types of infrasound events.

In summary, while there are significant differences in classification results across datasets, experiments using the same dataset with different classification methods demonstrate that our proposed method consistently achieves the best classification performance, thereby validating the effectiveness and reliability of our study.

4. In Table 6, please clarify how the 100% accuracy on LOTIS was computed—was this the average across folds, the best fold, or a separate hold-out test? The confusion matrices in Fig. 10 do not reflect the fold-averaging described.

Authors: Thank you very much for pointing out these issues. Our responses are as follows:

(1) Regarding the calculation of the 100% accuracy on the LOTIS dataset: According to the common data split method for this dataset adopted in the literature (Bryan et al., 2018; Zhao et al., 2024) (as shown in Table 1 for details), the test set contains a total of 103 signals, consisting of 34 AGW, 33 MAW, 17 MB, and 19 VE signals. The reported results are the averages obtained from multiple independent repeated experiments following this split scheme. It should be noted that since each experiment achieved completely correct classification (i.e., no misclassified samples), the average result across multiple independent repeated experiments remains 100%. Meanwhile, the Cohen's Kappa coefficient also reached 100%, further confirming that the agreement between the model's predictions and the ground-truth labels far exceeds the random level.

(2) Regarding the relationship between the confusion matrices in Fig. 10 and the above-mentioned fold averaging, we provide the following explanation. Figure 10 shows the classification results obtained on the LOTIS dataset using Branch 1 (GA-BiGRU), Branch 2 (MSCI-Net), and the full Infra-Net model, respectively. Again, following the common data split method from the literature (Bryan et al., 2018; Zhao et al., 2024), the test set contains 103 signals (34 AGW, 33 MAW, 17 MB, 19 VE). It is particularly important to note that, based on the multi-view learning strategy proposed in this paper, each original signal is transformed by the logarithmic wavelet scattering transform to extract three columns of scattering coefficients (corresponding to three different scattering paths). When only Branch 1 or Branch 2 is used for the experiment, since there is no confidence-based decision-making module to integrate the results from the multiple scattering paths, each scattering path is independently fed into the model and produces an independent classification result. Therefore, the number of samples involved in classification in Figs. 10(a) and 10(b) is three times the number of original signals, i.e., $103 \times 3 = 309$ samples. This explains why the sample counts in these two confusion matrices appear as multiples of 309. In contrast, when the full Infra-Net model is used, the confidence-based decision-making module performs local averaging and inner-product fusion on the probability outputs of the three scattering paths, ultimately outputting a single classification decision for each original signal. Hence, the

number of samples in Fig. 10(c) equals the size of the original test set, i.e., 103 samples.

In summary, the 100% accuracy reported in Table 6 is the average result of multiple independent repeated experiments under the current data split scheme, while the differences in sample counts among the confusion matrices in Fig. 10 arise from the number of scattering paths and whether the branch includes the fusion module.

5. Page 13, Line 325: Please provide visualization of the confidence-based fusion process with a supplementary figure showing how the per-column probabilities and the final inner product vary for a correctly classified vs. misclassified CTBTO sample.

Authors: Thank you very much for your suggestion regarding the visualization of the confidence-based fusion process. The working principle of the confidence-based decision-making module proposed in this paper is illustrated in Fig. 9. Specifically, a single infrasound signal is transformed by the logarithmic wavelet scattering transform to generate six columns of scattering features (corresponding to six different scattering paths). Each column of scattering features, treated as an independent view, is fed into two parallel branches (GA-BiGRU and MSCI-Net) for classification, so each branch obtains six sets of recognition probabilities. Subsequently, according to Equations (12) and (13), the six sets of probabilities within each branch are confidence-weighted averaged to obtain a comprehensive recognition probability for each branch regarding the signal. Finally, the inner product of the two branches' comprehensive probabilities is computed using Equation (14) to obtain the final joint recognition probability, and the event type corresponding to the larger probability value is taken as the final prediction.

To intuitively demonstrate the effectiveness of this fusion mechanism, we have supplemented the visualization of the probability change process as you requested. It should be noted that when both branches make correct predictions or both make incorrect predictions, the fusion result is consistent with the branch results and cannot reflect the decision-correcting capability of the fusion module. Therefore, we focus on representative samples where the two branches produce inconsistent classification conclusions. Figure S1 shows a sample where Branch 1 is correct and Branch 2 is incorrect; after inner-product fusion, the joint confidence favors the true class, and the final output is a correct recognition result. Figure S2 shows the opposite case (Branch 1 incorrect, Branch 2 correct); the fusion module similarly corrects the influence of the erroneous branch and

outputs a correct recognition result. These two cases fully demonstrate that the multi-view learning strategy combined with the confidence-based fusion module proposed in this paper can effectively leverage the complementarity of the two branches in feature perspectives, suppress the adverse effects of single-branch misjudgment, and thus significantly improve the accuracy and reliability of recognition results.

Once again, we thank the reviewer for your attention to visualization details. This supplement makes our explanation of the fusion mechanism more transparent and intuitive.

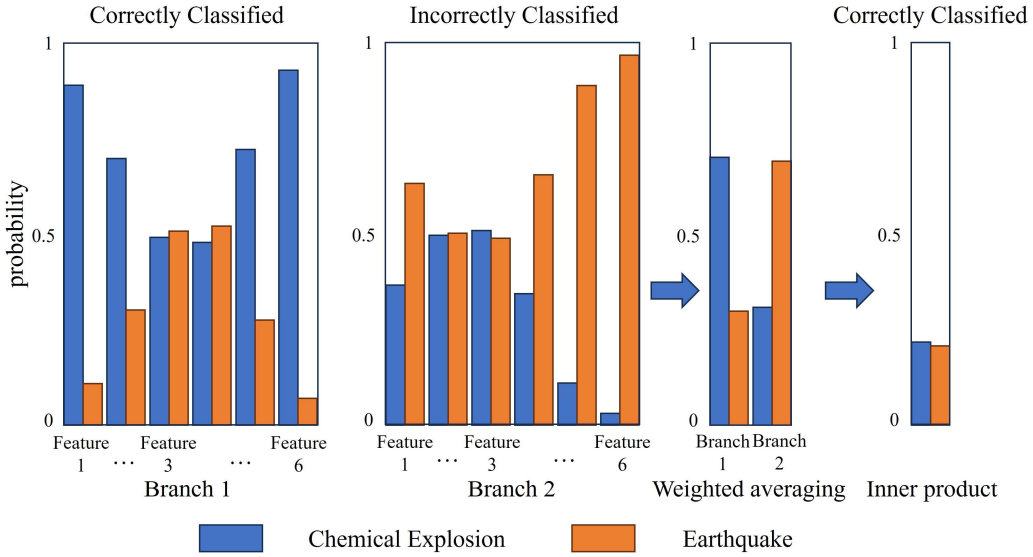

Fig. S1: Confidence fusion process when Branch 1 is correct and Branch 2 is incorrect.

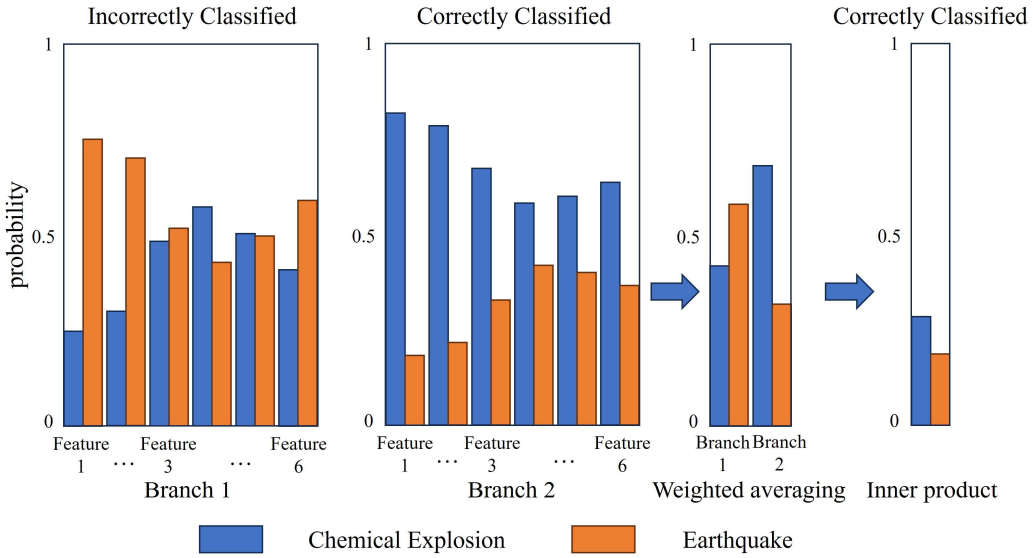

Fig. S2: Confidence fusion process when Branch 1 is incorrect and Branch 2 is correct.

The specific probability values are shown in the following two tables:

Branch 1		Branch 2	
Chemical Explosion	Earthquake	Chemical Explosion	Earthquake
0.891461	0.108539	0.367705	0.632295
0.697239	0.302761	0.496975	0.503025
0.49246	0.50754	0.509611	0.490389
0.479348	0.520652	0.346203	0.653797
0.723051	0.276949	0.10962	0.89038
0.929508	0.070492	0.0315	0.9685
0.70(Equation (12))	0.30(Equation (12))	0.31(Equation (13))	0.69(Equation (13))
Equation (14): 0.217(Chemical Explosion); 0.207(Earthquake)			

Branch 1		Branch 2	
Chemical Explosion	Earthquake	Chemical Explosion	Earthquake
0.2478933	0.75210673	0.81841689	0.18158311
0.30073237	0.69926763	0.78491426	0.21508574
0.48328087	0.51671916	0.67097843	0.32902154
0.57190692	0.42809314	0.58064693	0.41935307
0.50254619	0.49745381	0.59867698	0.40132302
0.4096241	0.59037584	0.6340121	0.36598784
0.42(Equation (12))	0.58(Equation (12))	0.68(Equation (13))	0.32(Equation (13))
Equation (14): 0.2856(Chemical Explosion); 0.1856(Earthquake)			

6. Suggest revising the final paragraph of the current Conclusion to include a brief statement on limitations and future domain adaptation work.

Authors: Thank you very much for your suggestion. We have added a description of the limitations of this method and future work on domain adaptation at the end of the conclusion, as follows:

In application efficacy, based on the parameter setting that "the optimal feature dimension satisfies scattering sampling frequency×time-invariant scale≈number of signal sampling points", this method is expected to achieve predictable and robust classification performance in similar infrasound monitoring scenarios. Despite the above results, this study still has several limitations that urgently need to be addressed in future work.

First, the performance gap between the public LOTIS dataset and the CTBTO field-measured data indicates that the current model's recognition capability is significantly constrained when dealing with signal distortions caused by extremely long-range propagation, and its cross-scene generalization ability still needs improvement. Furthermore, the ablation experiment results show

that the confidence-based fusion strategy proposed in this paper achieves only limited improvement over simple averaging fusion. To address these limitations, we will focus on the following research directions. First of all, we will introduce domain adaptation strategies, for example, reducing the model's dependence on specific monitoring environments through transfer learning to narrow the generalization gap. Second, we will collect and annotate more diverse infrasound event data (including various natural and anthropogenic events at different propagation distances and signal-to-noise ratios) to increase the coverage of training samples. Third, we will explore uncertainty-aware fusion methods to better handle conflicts among predictions from different views, thereby improving decision robustness and supporting more reliable natural hazard early warning.