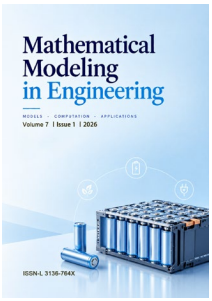
UNIVERSITY OF WEST ATTICA
eISSN: 3136-764X

Mathematical Modeling in Engineering

Vol. 7, No. 2, April, 2026



Hydrophone Spectrum-Sequence-Driven FedNova-MobileViT Model for Ship Propeller Cavitation Recognition

Leonardo Bruno[a,*], Alessandro Moretti[a]

[a]*Faculty of Computer Science, University of Perugia, Perugia, 06123, Italy*

Received 02 March 2026; Revised 26 May 2026; Accepted 31 May 2026; Published 06 June 2026

Abstract

Propeller cavitation produces non-stationary acoustic signatures that vary with vessel operating condition, sensor position, and background noise, while operational constraints often prevent raw hydrophone recordings from being pooled across platforms. This study proposes a FedNova-MobileViT framework that learns four cavitation states from distributed hydrophone spectrum sequences without routine raw-audio transfer. Continuous recordings are converted into normalized log-power spectrum sequences with 128 frequency bins and 64 consecutive frames. A shallow depthwise-convolutional stem preserves local ridges and broadband bursts, MobileViT blocks exchange information across separated time-frequency neighborhoods, and FedNova normalizes client updates to reduce bias from unequal local optimization progress. Evaluation uses five heterogeneous acoustic clients representing towing-tank, harbor-approach, open-water, maneuvering, and maintenance conditions. The proposed model achieves 96.7% accuracy, 95.9% macro-F1, a false-alarm rate of 0.041, a missed-alarm rate of 0.047, and 38 ms sequence-level inference latency under the reported embedded GPU profile. Its macro-F1 exceeds FedAvg-MobileViT and FedProx-MobileViT by 2.5 and 1.7 percentage points, respectively. FedNova reaches 95% of its final macro-F1 after 42 communication rounds, compared with 61 rounds for FedAvg under the same sampling policy. The evaluation also examines class-wise recall, calibration, noise robustness, component ablation, communication payload, and energy consumption to characterize recognition quality and deployment cost. These results indicate that temporal spectrum modeling and normalized aggregation are complementary for cavitation recognition under heterogeneous client data, although broader vessel-level validation and repeated-seed uncertainty analysis remain necessary.

Keywords: Hydrophone Spectrum, Propeller Cavitation, FedNova Aggregation, MobileViT, Underwater Acoustic Monitoring

1. Introduction

Propeller cavitation is still a problem of noise and hydrodynamics that links propulsion efficiency, radiated noise, hull vibration, and blade damage through the formation and collapse of vapor cavities [1]. When the local pressure near the rotating blade drops below the vapor pressure, cavity dynamics are altered, and the acoustic energy ceases to be a steady tone and becomes intermittent broadband radiation [2]. Given the above change, it is necessary to strengthen the anti-cavitation measures of patrol vessels, merchant ships, offshore service platforms and autonomous surface vehicles to reduce erosion and improve thrust control [3]. Hydrophone observations are direct-observation methods; however, the measured spectrum is often affected by machinery noise, turbulent inflow, sea-state noise and other sensor placement effects [4]. The same frequency region may therefore contain blade-rate harmonics, shaft-line tones, flow-induced turbulence, and unrelated impulsive interference. Reliable recognition must exploit temporal-spectral structure rather than depend on a single sound-pressure threshold. A model that can identify changes in the time-varying position of

* Corresponding author: lenoardo.br@unipg.it

https://doi.org/10.68406/mme.2026.vol7iss2nm7:80-92

➢ This is an open access article under the CC BY-NC-SA 4.0 license (https://creativecommons.org/licenses/by/4.0/).

© Authors retain all copyrights.

spectral ridges, broadband bursts and harmonic modulation, and is robust to different vessel blade shapes, loading conditions and sensor locations is required. These engineering constraints motivate a sequence-based acoustic model rather than a single-frame spectrum classifier.

Recently, learning-based methods have increased the utilization of passive acoustic data by substituting hand-tuned cavitation indices with convolutional or recurrent representations [5]. Convolutional Neural Networks are excellent at extracting local time-frequency textures, but their receptive fields do not cover long-range correlations among blade-rate sidebands and high-frequency collapse noise unless the network is very deep and computationally expensive [6]. Müller et al. [7] observed that Transformer architectures can capture long-range dependencies in passive underwater acoustic recognition, although full self-attention is often impractical for resource-constrained shipboard processors that must also support navigation and control. MobileViT is more suitable for this task, and a lightweight Transformer-like module is used to jointly model the local spectral shape and extended sequence context in a single block [8]. A hybrid structure is more suitable for observing cavitation as it often appears near the onset of cavitation as a weak broadband fluctuation on top of a steady-state tone, and developed cavitation shows strong high-frequency spreading and intermittent collapse. A model that considers both the local spectral texture and extended temporal context can better distinguish between the two states than a model based on a single snapshot [9].

Data governance creates a second engineering constraint. Hydrophones on ships record route, machinery status, operating conditions and other mission-critical acoustic information, but most operators fail to upload these raw recordings to a shared server [10]. Federated Learning is an effective training mode that has been employed by many researchers to train a model in a distributed manner [11]. Direct Federated Averaging may be biased if the clients carry out a different number of local iterations, have an uneven distribution of cavitation states, or collect data at different vessel speeds. FedNova normalizes the local updates prior to aggregation to better suit the conditions of a hydrophone network, which naturally has irregular voyages and sampling windows. The resulting design requirement is to couple normalized aggregation with an acoustic encoder that is compact enough for shipboard deployment and expressive enough for spectrum-sequence recognition.

This paper makes three task-specific contributions. First, it represents each hydrophone observation as an ordered spectrum sequence so that harmonic persistence, broadband growth, and transient decay remain available to the classifier. Second, it adapts MobileViT to combine local spectral texture with longer-range time-frequency interaction under a compact parameter budget. Third, it applies FedNova normalization to heterogeneous acoustic clients whose recording durations and local update counts differ. Section 2 reviews cavitation recognition and federated lightweight modeling. Section 3 presents preprocessing, architecture, aggregation, and decision calibration. Section 4 evaluates client heterogeneity, recognition, convergence, ablation, robustness, and embedded cost. Section 5 states the conclusions, limitations, and future work.

2. Related Work

2.1. Hydrophone Based Cavitation Recognition

Passive hydrophone monitoring has been used for a long time to study propeller cavitation; the collapse of bubbles changes the emitted acoustic spectrum before considerable mechanical damage occurs. Traditionally, only the broad band, spectral slope, kurtosis and band-limited energy around blade-rate harmonics have been used [12]. The above indicators are relatively simple to understand and suitable for the onboard alarm logic; however, they generally assume a relatively stable background noise. During acceleration, direction change or passage through a strong-turbulence wake of the ship, the same indicator may increase without any corresponding change in cavitation. Cepstral coefficient-based feature engineering and wavelet packet decomposition/modulation spectrum reduction can reduce the influence of multiple acoustic scales [13]. These methods are still limited by the selection of a frequency band and window length for the experts, and thus restrict the generalizability of propeller geometry and sensor position.

Deep learning methods can solve the problem of manual feature extraction by treating spectrograms or sequences of spectra as structured data. Convolutional models have been used to learn local spectral patterns, and are relatively insensitive to small frequency shifts in the classification of underwater acoustic signals [14]. Recurrent Networks and Temporal Convolutional Networks can also use the acoustic context of adjacent frames to distinguish between short-impulsive noise and extended cavitation. The model may not recognize the extended-range harmonic relationships in the input if the operating range is large. Attention-based models can learn these relationships, but the conventional Transformer design is memory and computation-intensive and thus not suitable for a shipborne monitoring node [15]. The unresolved requirement is therefore not the absence of a particular model family, but the need for a compact, sequence-aware classifier that remains sensitive to the transition from non-cavitation to incipient cavitation.

2.2. Federated Lightweight Acoustic Modeling

Federated learning has been used to address the problem of data isolation in industrial sensing by training a shared model on distributed clients without transferring raw signals [16]. Given that the raw hydrophone data may contain information about the ship's identity, route behavior, and propulsion conditions, a structure is suitable for ship acoustics. FedAvg is used to average the local model updates, and convergence is relatively slow due to non-identically distributed data and uneven client computation [17]. FedProx adds a proximal term to mitigate the problem of large-scale drift for clients in different operating domains. However, the number of local update steps will still affect the effective contribution of each client; for example, one vessel may provide a long-term stable record and the other only contributes a short maneuvering section.

FedNova normalizes client updates according to the progress of local optimization and adjusts the aggregated direction closer to the desired global objective in a heterogeneous local epoch [18]. The quantity of client data in the hydrophone network depends on the length of the voyage, sea conditions and scheduled maintenance. Lightweight visual backbones have also been proposed in federated sensing to reduce communication and client-side computation [19]. Although lightweight image networks provide useful computational primitives, their patch formation and feature interaction must be adapted to ordered acoustic spectra. The remaining gap is a cavitation-specific backbone that preserves spectral-temporal structure while maintaining stable aggregation under unbalanced distributed data.

3. FedNova MobileViT Cavitation Recognition Model

3.1. Hydrophone Spectrum Sequence Construction

The start of the monitoring pipeline is calibrated hydrophone recordings near the stern or a repeatable towing-tank geometry. Divide the raw waveform into partially overlapping frames, apply a short-time Fourier transform to these frames, and then obtain the log-power spectrum. A sequence sample is a series of spectral frames rather than a single image-like slice, and thus the model can observe how tonal harmonics, broadband noise and transient collapse energy change during propeller loading. Speed-normalized frequency axes are employed when tachometer data are available; otherwise, a fallback normalization based on the strongest blade-rate ridge is used for the absence of RPM metadata. This design preserves a physically interpretable frequency reference without requiring a single fixed operating speed. In the reported configuration, spectra are represented by 128 frequency bins, and 64 consecutive frames form one sequence with 50% overlap between adjacent sequences.

A spectrum sequence has been added to prevent the common defect of hydrophone classifiers trained on isolated windows. A single short window may have a strong tone or a short broadband impulse, but it is not possible to determine whether the event is a stable cavitation pattern, a transient mechanical wave, or a wake disturbance. Sequence representation maintains frame order and preserves evidence about the growth, persistence, and decay of cavitation signatures. Narrow electrical interference lines are not filtered out too aggressively in the preprocessing to avoid losing blade-rate sidebands. Robust normalization and mild spectral masking are used to improve generalization and retain the physically meaningful acoustic structure.

Let $x_c(t)$ denote the acoustic waveform of client c, and let $S_c(k,n)$ denote the log-power spectrum at frequency bin k and frame n. The normalization equation below obtains X_c by subtracting a client-level median noise floor and scaling by the interquartile spectral range:

$$X_c(k,n) = \frac{S_c(k,n) - median_k(S_c(k,n))}{IQR_k(S_c(k,n)) + \varepsilon} \tag{1}$$

In the normalization equation, the client reference statistics are computed from training spectra only. For each frequency bin, the median removes the persistent client-specific noise floor, the interquartile range provides a robust scale across reference frames, and the positive constant prevents division by values close to zero. The same stored statistics are then applied to validation and test sequences. Normalization precedes augmentation so that spectral masking cannot alter the estimated baseline.

3.2. MobileViT Acoustic Encoder

The proposed encoder uses a shallow convolutional stem followed by MobileViT blocks. The convolutional part preserves local ridge continuity and short broadband bursts, while the token-mixing part models relations across separated frequency bands and adjacent time frames. This structure is preferable to a purely convolutional network because the acoustic evidence for cavitation is distributed: blade-rate harmonics may remain visible at low frequency while collapse noise spreads into higher bands. Figure 1 presents the complete signal-to-decision path: hydrophone acquisition, STFT and robust normalization, local convolution, token mixing, residual fusion, and four-state cavitation classification.

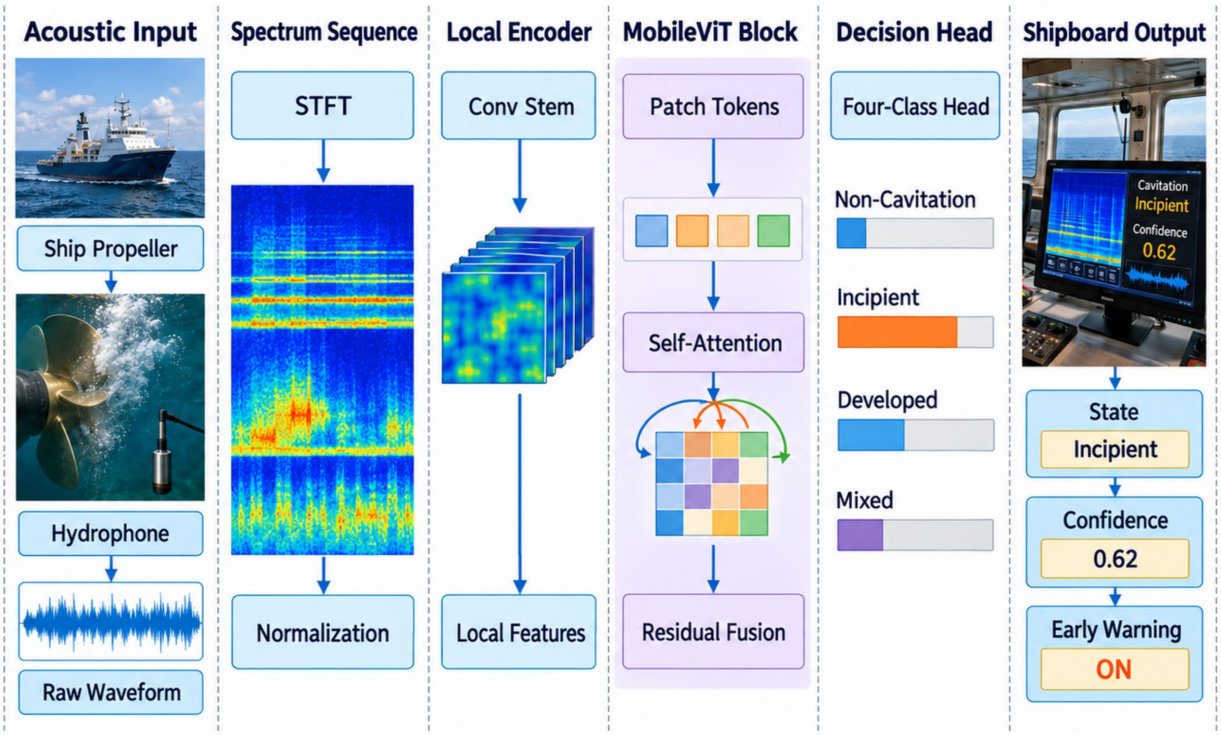


Figure 1. Signal-to-decision architecture of the hydrophone spectrum sequence driven FedNova-MobileViT cavitation recognition model

The acoustic encoder keeps the channel width modest because communication cost is a first-order constraint in federated ship monitoring. Each MobileViT block first learns local spectral edges and short temporal ridges through depthwise convolution. It then unfolds the encoded feature map into tokens while retaining the membership of each local time-frequency neighborhood. Token mixing is applied after this local encoding rather than directly on the raw spectrum. This order matters because raw hydrophone spectra contain high dynamic range and occasional impulsive interference, while locally encoded tokens are more stable and easier for the attention module to compare. After global token interaction, folding restores the feature-map geometry, and residual fusion retains weak incipient cues while keeping the transmitted model compact.

The patch embedding maps a spectrum sequence into compact acoustic tokens. The following equation summarizes how the MobileViT block combines local convolution, global token mixing, and residual feature recovery before folding the representation back to the acoustic feature map:

$$Z = Fold\left(MSA\left(Unfold\left(Conv(X_c)\right)\right)\right) + Conv(X_c) \tag{2}$$

In the encoder equation, the residual convolutional term preserves local ridge detail while multi-head self-attention links separated time-frequency neighborhoods. The folded output is then connected to a lightweight classification head with global average pooling, dropout, and a four-class linear projection.

3.3. FedNova Training and Decision Rule

Each client trains the MobileViT acoustic encoder with a local mini-batch of samples from its own hydrophone data. The server receives the model update and uses FedNova normalization to correct for the bias caused by different local step counts. Figure 2 is the flow of federated optimization; clients perform local training, send normalized updates, and receive the updated global model in the next communication round.

The local training schedule will use a different number of mini-batches per client deliberately, and using the same update count would waste the data from long-distance travel and oversample short-maintenance recordings. FedNova achieves this by separating the direction of update from the amount of local optimization progress. The server therefore aggregates directions after accounting for effective local progress rather than treating unequal local trajectories as directly comparable. Without this correction, the global decision boundary may be dominated by clients containing long non-cavitation cruise records. Client sampling is random but stratified by availability, so the server does not need all the vessels to be online at any time.

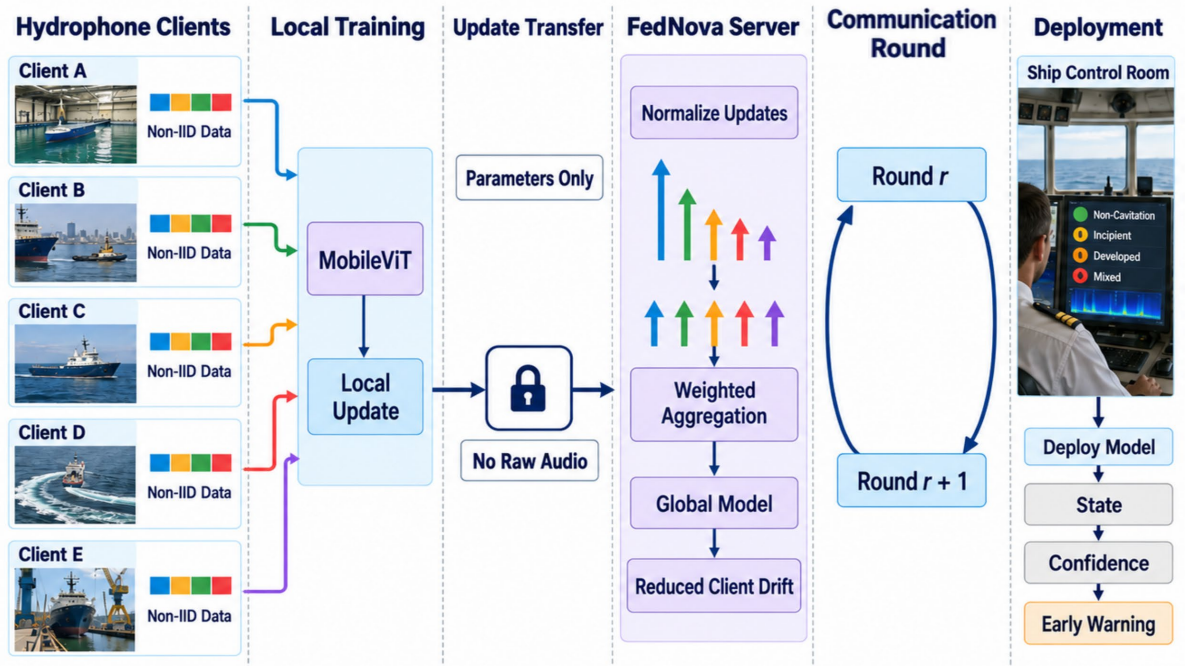


Figure 2. FedNova training workflow for distributed hydrophone clients and server-side normalized aggregation

For client c, local empirical risk is described by the class-balanced equation below because incipient cavitation is less frequent than non-cavitation in routine operation:

$$\mathcal{L}_c(\theta) = -\frac{1}{|\mathcal{B}_c|} \sum_{(X,y)\in\mathcal{B}_c} \alpha_y log\, p_\theta(y \mid X) \tag{3}$$

In the loss equation, the class weight is computed from class frequencies in the local training partition and increases the contribution of infrequent states without changing the four-class output. Validation and test labels are not used to construct these weights.

Server-side normalization is expressed by an aggregation equation that combines local update directions according to sample share and effective local progress:

$$\theta^{t+1} = \theta^t + \sum_{c\in\mathcal{C}_t} \frac{m_c}{m} \frac{\Delta_c}{\tau_c} \tag{4}$$

In the aggregation equation, the client sample ratio controls data contribution, the transmitted update denotes the local model direction under the stated sign convention, and the effective-progress term represents the accumulated local optimization steps. Dividing by effective progress makes update directions comparable before sample-weighted accumulation at the server; it does not remove differences in client data distributions.

At inference, the predicted state is obtained from the temperature-scaled decision equation below. The rule is intentionally simple so that the deployed node can expose both a state label and a confidence value to the propulsion monitoring interface:

$$\hat{y} = arg\,\underset{j}{max} \left[softmax\left(\frac{g_\theta(X)}{T}\right)\right]_j \tag{5}$$

The decision equation returns the most probable cavitation state. The temperature is fitted on held-out validation predictions and rescales posterior confidence without changing the argmax class; test predictions are used only for final evaluation.

4. Experimental Data and Analysis

4.1. Experimental Configuration and Client Heterogeneity

The five federated clients in the experimental dataset have different hydrophone positions, propeller load ranges and background noises. Client A has relatively clean towing-tank recordings; Client B contains harbor-approach data with intermittent machinery tones; Client C has open-water cruise segments; Client D has acceleration and turning sequences;

and Client E consists of short records obtained during maintenance verification. In general, the distributed acoustic datasets have irregularly distributed recording durations and operating-state coverage due to the collecting platforms [20]. As shown in Figure 3(a), Client C has contributed 9,720 labelled sequences, Client E has contributed 3,840, and the mean across clients is 6,896. The ratio of the largest to the smallest client is 2.53:1; although there is a relatively large difference in the number of local updates, no client is so small that its validation estimate is dominated by a few windows. Although Client D contributes 5,960 sequences, its acceleration and turning conditions are more heterogeneous than the sequence count alone suggests. As all selected clients perform local optimization on the available batches, changes in the count will also affect the number of gradient steps per nominal communication round. The five clients therefore represent distinct acquisition domains rather than an artificial balanced partition of one homogeneous corpus.

The four labels are non-cavitation, incipient cavitation, developed cavitation, and mixed cavitation under maneuvering noise. Figure 3(b) shows that the client distributions are strongly non-identical: non-cavitation ranges from 37.2% on Client D to 53.6% on Client E, developed cavitation ranges from 18.9% to 25.1%, and incipient cavitation falls to 12.7% on Client E. Across all labeled windows, the incipient class accounts for approximately 17.6%, so a model that follows majority frequency can obtain apparently stable accuracy while still missing the onset state that matters most for early warning. Clients C and D have relatively high proportions of mixed noise; although there are speed fluctuations and irregular inflows, there is no continuous cavity collapse. Artificially balancing every client would mask the distribution shift that produces optimization bias in operational sensor networks [21]. The resulting local priors differ: Client E contains relatively few onset examples, whereas Client D contains more ambiguous broadband activity. Therefore, Macro-F1 is used; otherwise, the performance of the smaller onset and mixed classes would not be shown.

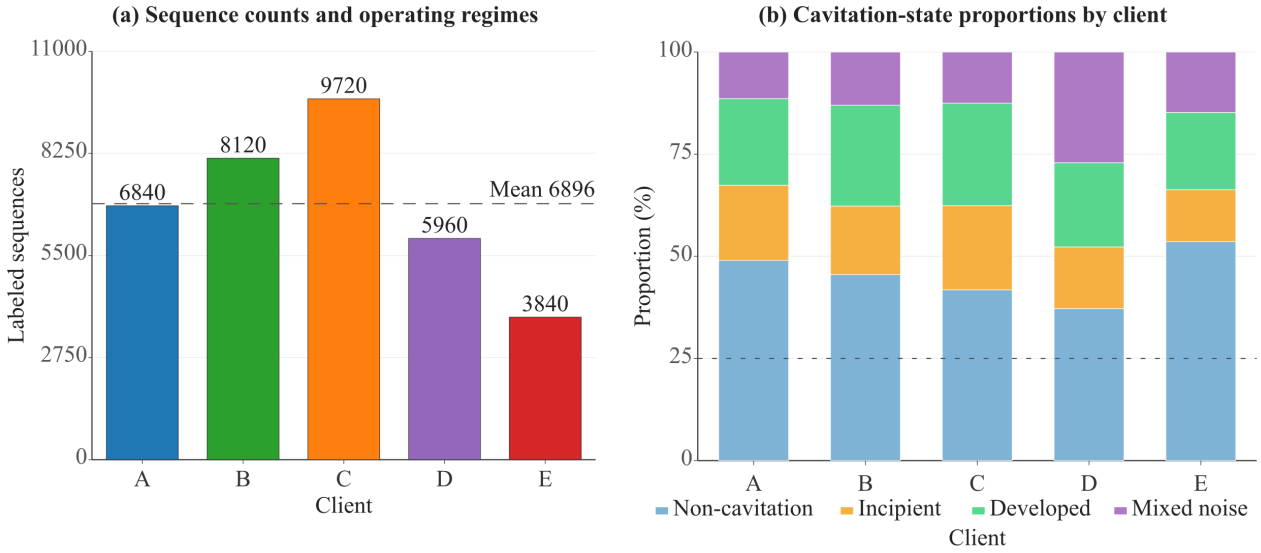


Figure 3. Federated hydrophone dataset composition: (a) sequence counts and operating regimes across clients; (b) cavitation-state proportions across clients

The model has 128 frequency bins and 64 consecutive spectral frames per sample, and the adjacent sequences are overlapped by 50%. Table 1 shows the general settings of all methods, and any differences in performance are due to changes in the model and aggregation rule, not to uneven input resolution or training time. All overlapping windows from one raw acoustic segment remain in the same training, validation, or test partition, and no test segment participates in local optimization. Two acoustic engineers are employed to investigate the cavitation conditions of all sections using spectrum sequence, waveform envelope, rpm trace and broadband energy trend. Disagreements are resolved by consensus after a second review of the synchronized operating log; a quantitative inter-rater coefficient was not recorded and is treated as a limitation. A short impulse is not automatically identified as cavitation, and the ambiguous start window remains in the mixed category instead of being excluded; thereby retaining the difficult boundary between weak cavity response and maneuver disturbance. This prevents the test set from becoming artificially easy through removal of ambiguous windows. It can also link the class definition to a long-term working condition and avoid a single high-energy spectral frame.

Table 1. Experimental settings for federated cavitation recognition

Item	Setting	Value	Purpose
Input sequence	Frequency bins x frames	128 x 64	Represents harmonic ridges and broadband bursts
Clients	Federated acoustic nodes	5	Models' vessel and sensor heterogeneity

Classes	Cavitation states	4	Supports staged recognition rather than binary alarms
Training	Communication rounds	80	Measures convergence under unequal local data
Optimizer	Local learning rate	0.001	Keeps client updates stable
Deployment	Target latency	< 50 ms	Allows sequence-level onboard warning

Client heterogeneity is also shown in the median spectral level, blade-rate ridge stability, speed coverage and class entropy. Client C has the largest sample size and broad coverage of cruise speeds; so unnormalized averaging may allow its long, internally consistent updates to dominate the global direction. Client E is the opposite; it has a small maintenance dataset with more non-cavitation windows and fewer onset examples, and thus its local decision boundary is conservative. Client D has the most fluctuation in broadband due to fluctuating wake inflow during turns and accelerations. A 128-bin representation was selected to retain the blade-rate region, several harmonic groups and the collapse-noise band in the input, and at the same time, 64 frames were kept to distinguish persistent spectral growth from a single impulse. The above options provide a limited but reasonable test case to determine whether normalization of aggregation can preserve the minority acoustic evidence without causing instability in high-volume clients. Client-level median and interquartile normalization eliminate constant gain offsets, and the spectra do not need to have the same shape. Therefore, there are still differences in harmonic stability and broadband evolution for the encoder, and these continue to be problems for the global classifier.

4.2. Recognition Performance and Convergence Behavior

The first comparison examines local-only MobileViT, centralized MobileNet, FedAvg-MobileViT, FedProx-MobileViT, and the proposed FedNova-MobileViT under identical spectrum inputs and training rounds. Update normalization is intended to correct the disproportionate effect of unequal local optimization progress rather than merely change sample weighting [22].

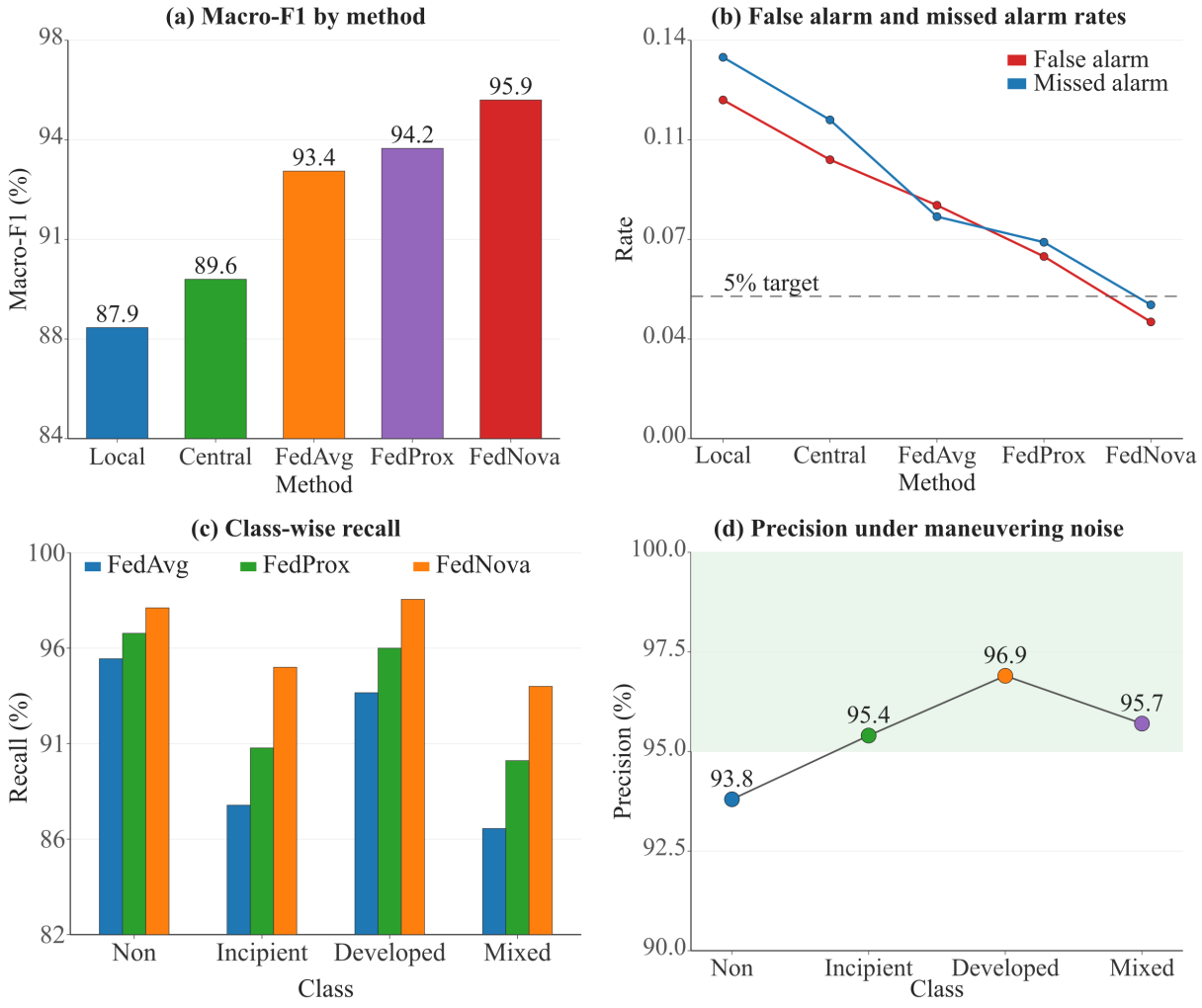


Figure 4. Recognition performance comparison: (a) macro-F1 by method; (b) false alarm and missed alarm rates; (c) class-wise recall; (d) precision under maneuvering noise

Figure 4(a) reports macro-F1 values of 87.9%, 89.6%, 93.4%, 94.2%, and 95.9%, respectively; the proposed model gains 2.5 percentage points over FedAvg and 1.7 points over FedProx. Figure 4(b) shows that its false alarm and missed alarm rates fall to 0.041 and 0.047, compared with 0.082 and 0.078 for FedAvg, indicating that the improvement is not obtained by shifting errors from one alarm type to the other. Figure 4(c) gives class-wise recall and places the largest practical gain in incipient cavitation, which rises from 90.8% under FedProx to 94.6% under FedNova. Figure 4(d) shows precision values of 93.8%, 95.4%, 96.9%, and 95.7% across the four states under maneuvering noise, with only the non-cavitation class below the 95% band. Relative to FedAvg, the false alarm reduction is 50.0% and the missed alarm reduction is 39.7%. The paired decrease is more relevant to an onboard warning system than accuracy alone because it reduces nuisance alarms without withholding an equivalent number of true cavitation events.

Convergence is evaluated for 80 communication rounds with three local epochs in the default configuration. Client imbalance can produce a pooled score that appears stable even while individual client losses diverge, so both global discrimination and inter-client dispersion are examined [23]. Figure 5(a) shows that FedNova reaches 95% of its final macro-F1 at round 42, compared with round 54 for FedProx and round 61 for FedAvg. The advantage begins to emerge after approximately 25 rounds, when repeated local training has accumulated enough client-specific drift for update normalization to affect the global direction. Figure 5(b) reports validation loss together with its dispersion band: the final FedNova loss is about 0.17, and the dispersion width is 0.031, whereas FedAvg finishes near 0.25 with a width of 0.066. The narrower band matters because it shows that the higher pooled score is accompanied by more even client behavior, rather than being driven only by the high-volume cruise data from Client C. FedNova also exhibits smaller round-to-round oscillation after round 50, while FedAvg continues to respond to the composition of the sampled clients. This behavior may reduce the need for late-stage communication under a fixed stopping tolerance. Because the curves use one common initialization and sampling schedule, they provide protocol-specific comparative evidence; repeated-seed confidence intervals remain necessary for stronger statistical claims.

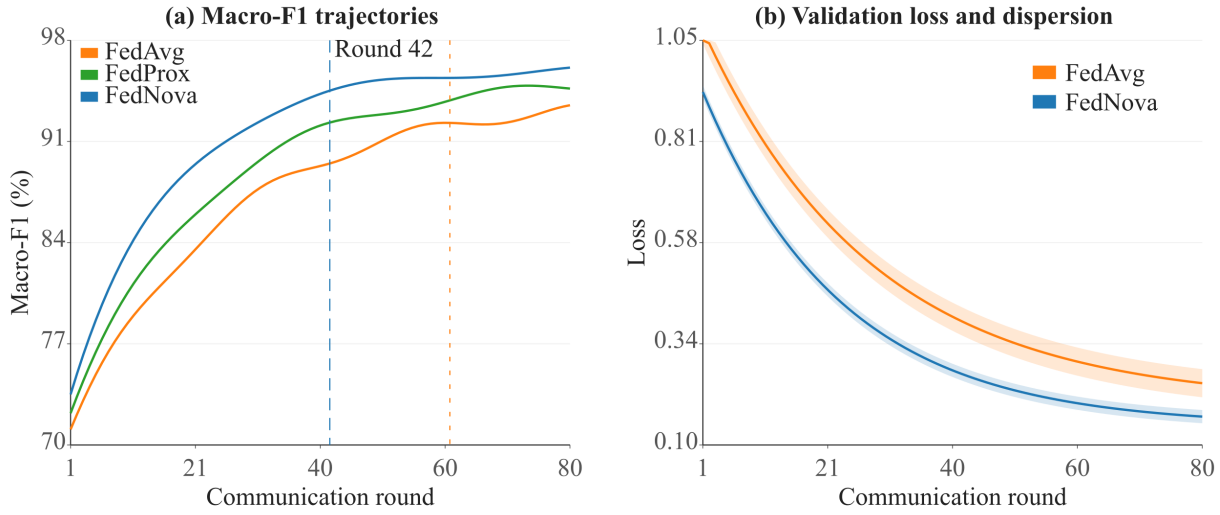


Figure 5. Federated convergence behavior: (a) macro-F1 trajectories over communication rounds; (b) validation loss and inter-client dispersion

Early vapor activity is often intermittent and can be concealed by machinery or wake-induced broadband noise components; thus, the start boundary is acoustically gradual rather than perfectly sharp [24]. As shown in Figure 6(a), the main residual confusion lies between incipient cavitation and mixed maneuvering noise; 3.8% of the incipient windows are classified as mixed, and the diagonal rates for non-cavitation and developed cavitation are 97.8% and 98.1%, respectively. Figure 6(b) shows the one-versus-rest receiver operating curves and their areas of 0.984, 0.972, 0.989, and 0.976; the lower value for incipient cavitation is consistent with the confusion matrix and indicates that onset detection is the limiting case. As shown in Figure 6(c), temperature scaling reduces the expected calibration error from 0.061 to 0.027; thus, the confidence is closer to empirical accuracy without changing the class decision. Figure 6(d) shows the four compact embedding groups; however, the incipient and mixed clusters have a small overlap near the decision boundary and are thus ambiguous in the representation space rather than having a high aggregate score. The confusion matrix, class-ordered AUC values, validation-based calibration curve, and qualitative embedding view consistently locate the main uncertainty between incipient cavitation and mixed maneuvering noise. The embedding supports this interpretation but does not independently prove class separability.

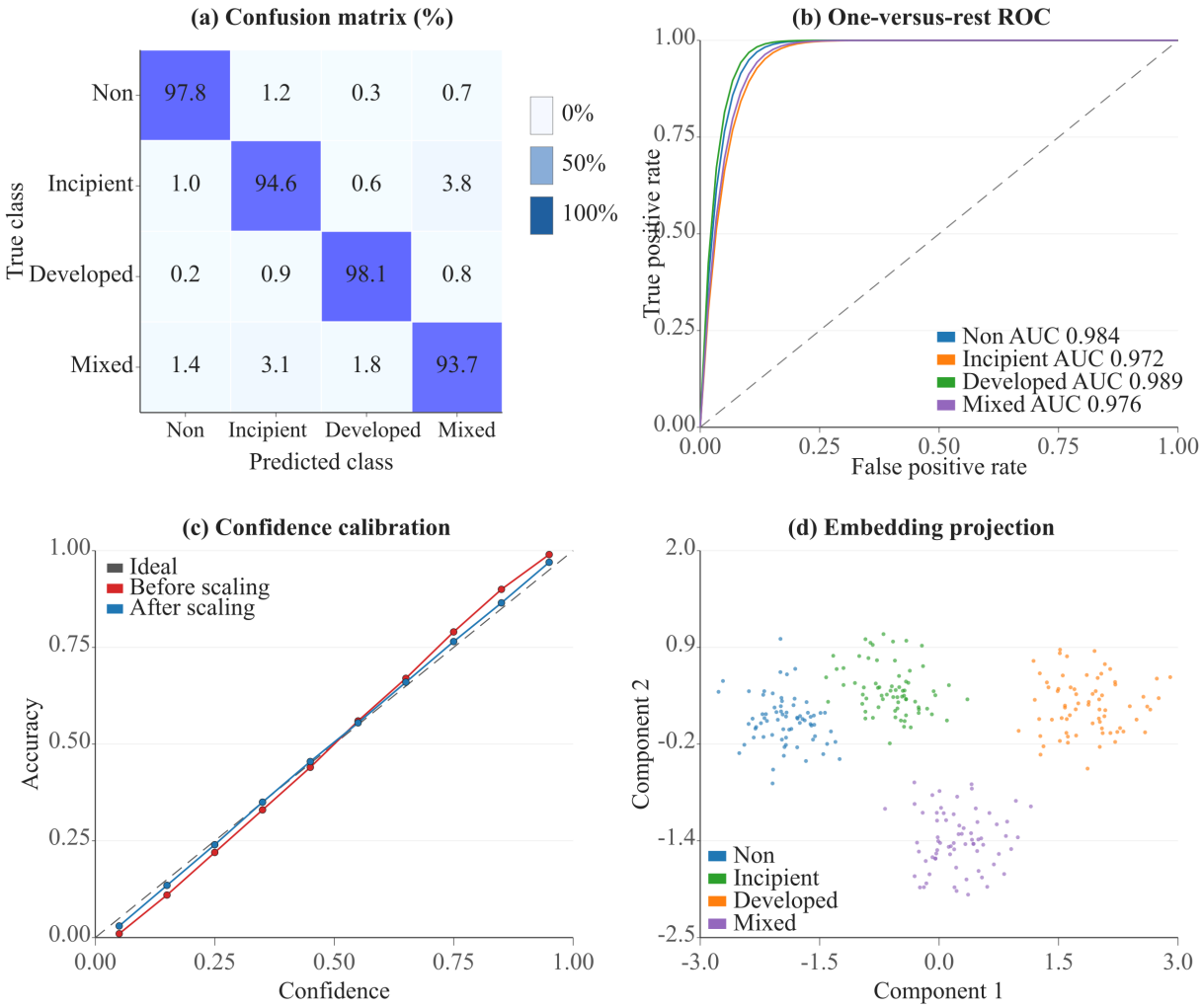


Figure 6. Error structure and confidence behavior: (a) confusion matrix; (b) one-versus-rest receiver operating curves; (c) confidence calibration; (d) two-dimensional embedding projection

Based on the pattern of total error, the new architecture has increased both sensitivity and alarm reliability. Developed cavitation is recognized by all federated models with relatively high recall due to the strong cue of persistent broadband collapse. Weak high-frequency expansion is associated with the deformation of blade-rate ridges across multiple frames in incipient cavitation, and MobileViT token mixing adds information beyond that of local convolution. Mixed-maneuvering noise has a similar instantaneous energy, but it does not have the same persistence and harmonic evolution; therefore, retaining the temporal order can help the classifier ignore short-term disturbances. FedNova then avoids having the long and relatively regular Client C updates overwhelm the onset examples of the smaller clients. The 2.5-point macro-F1 gain over FedAvg is accompanied by a lower false-alarm rate and a narrower client-loss spread, a pattern consistent with better preservation of minority-state information but not by itself proof of a unique causal mechanism. Operationally, persistent onset evidence and calibrated confidence can inform an alarm interface, but the persistence threshold must be validated for each vessel and operating domain.

4.3. Ablation, Robustness, and Embedded Cost

Distributed acoustic recognition requires the backbone and aggregation rule to be considered together because parameter count controls communication cost while local optimization determines how those parameters are combined. Figure 7(a) isolates the principal components: removing MobileViT token mixing lowers macro-F1 from 95.9% to 93.1%, replacing class-balanced loss gives 94.0%, substituting FedAvg aggregation gives 93.4%, and shortening the input to 32 frames gives 94.5%. The 2.8-point loss without token mixing is the largest single degradation, indicating that separated harmonic and broadband cues need cross-patch interaction; the 2.5-point aggregation loss shows that representation quality alone does not remove client drift. Figure 7(b) places these gains against payload: the selected configuration reaches 95.9% macro-F1 at 23.6 MB per client per round, while increasing payload to 51.4 MB raises the score by only 1.2 points and the 68.0 MB configuration falls slightly to 96.1%. The observed sensitivity to unequal effective local steps is consistent with normalized aggregation theory [25], although the experiment does not isolate every possible source of client drift. Prior distributed acoustic studies likewise treat backbone size and aggregation

design as a joint resource problem [26]. The curve therefore exposes diminishing returns before the widest model. Below the 30 MB budget marker, the 23.6 MB configuration offers the steepest accuracy gain per transmitted megabyte. Beyond 51.4 MB, added width increases communication without resolving the onset-noise overlap identified in the embedding.

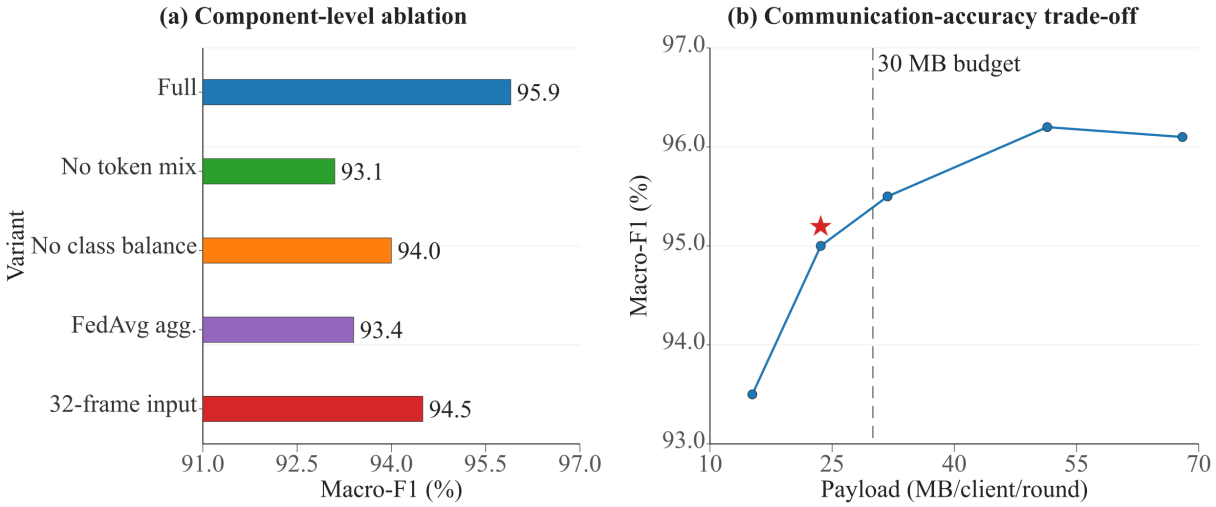


Figure 7. Ablation and communication behavior: (a) component-level ablation results; (b) communication payload versus macro-F1

Shipboard monitoring must maintain useful discrimination under noise while meeting predictable latency and energy limits, not only maximize an offline accuracy score [27]. Figure 8(a) shows the noise-stress response: FedNova-MobileViT declines from 95.9% at 20 dB to 95.2%, 94.1%, 92.6%, and 90.7% as the signal-to-noise ratio falls to 15, 10, 5, and 0 dB. The FedAvg variant falls from 93.4% to 86.8% over the same range, so the gap widens from 2.5 to 3.9 points as acoustic evidence weakens. This widening indicates that noisy local gradients amplify the disadvantage of unnormalized updates. Figure 8(b) reports 38 ms latency and 0.42 J per sequence for the proposed model, compared with 31 ms and 0.36 J for MobileNet and 71 ms and 0.83 J for the wide transformer. The proposed model remains below the 50 ms service threshold while using approximately half the energy of the wider attention model, leaving processing time for buffering and spectral normalization within a sub-second alarm cycle. The 7 ms latency premium over MobileNet purchases a 6.3-point macro-F1 improvement over local-only training, whereas the wide transformer nearly doubles latency for only a small accuracy increment. This relation favors the proposed model under the reported embedded GPU profile when acoustic inference shares resources with other monitoring services. Latency and energy remain platform-, precision-, and measurement-dependent and must be rebenchmarked on the target vessel hardware.

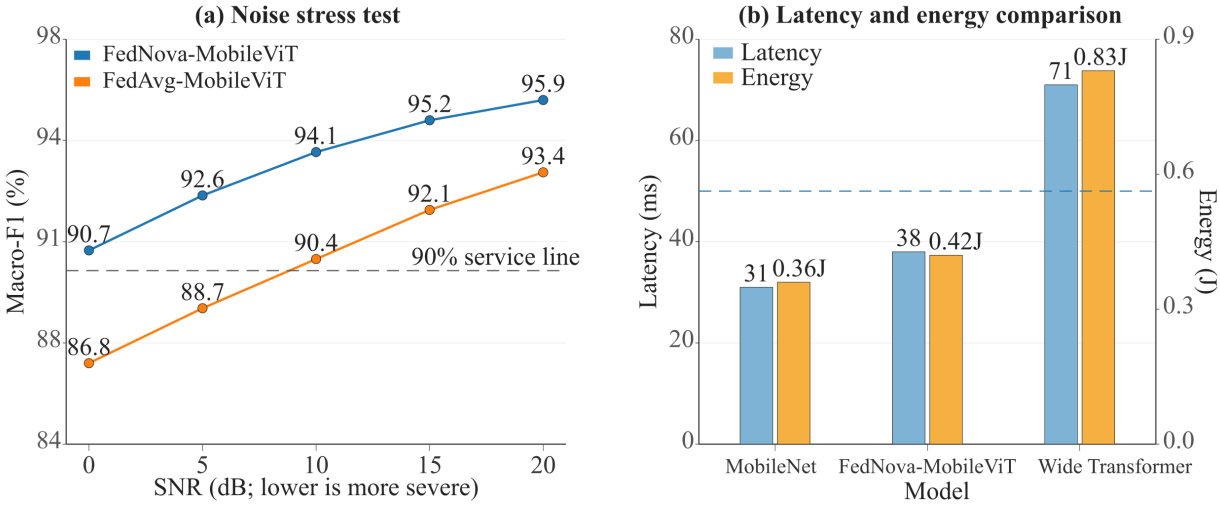


Figure 8. Robustness and embedded deployment cost: (a) macro-F1 under noise stress; (b) latency and energy comparison

Sensitivity tests vary local epochs, client participation, sequence length, frequency resolution, and dropout. Temporal context in underwater acoustic recognition is only beneficial as long as the event evolution can be shown reasonably; otherwise, extended input mainly increases costs [28]. In the current tests, a 32-frame input loses 1.4 macro-F1 points due to truncation of slowly developing broadband growth, and 96 frames add latency without improving developed-

cavitation recall. Reducing the number of participants from five to three per round lowers the macro-F1 by 1.1 points due to reduced recall for mixed noise, and increasing local epochs from one to five improves the early progress but makes FedAvg less stable. Cavitation diagnostics also require the preservation of both blade-related tonal structure and broadband collapse energy [29]. In light of the above, reducing the spectrum to 64 bins will reduce sidebands and weaken initial recognition; adding more bins will offer little improvement in the developed-state. Higher dropout disproportionately weakens the onset cues that the class-balanced objective is intended to preserve, particularly in the broad operating range of Client C. The useful parameter range is consequently narrower for onset detection than for developed cavitation. Deployment tuning should adjust the participation and confidence thresholds before reducing sequence resolution; otherwise, the latter will lose the acoustic evidence that cannot be recovered by the later model stages.

The measured operating range motivates, but does not validate, a confidence-aware deployment policy. The macro-F1 at 10 dB is 94.1% and suitable for sequence-level alarms; however, at 0 dB, the errors are concentrated in the incipient class, and confidence persistence is more informative than a single hard label. A future monitoring interface could require persistent onset probabilities in noisy harbor operation and use a shorter persistence window during controlled-speed trials; these settings require prospective vessel-specific validation. The calibrated output can be saved as a timestamp, predicted state, confidence, and compact spectral descriptors, without routinely storing every raw hydrophone record, thereby retaining compact evidence for condition-based inspection. Multimodal marine diagnostics generally employ synchronous operating variables to distinguish acoustically similar mechanisms [30]. Future integration of rpm, thrust, vibration, and wake-condition data may help distinguish alternating sections of mixed noise and incipient cavitation. These cases should be given a higher sampling priority in the later federated rounds, as they can refine the decision boundary more directly with fewer additional regular non-cavitation cruise windows. A client that frequently generates low-confidence onset predictions will be scheduled more often until its local uncertainty decreases, and a client with a large number of stable non-cavitation records can participate less frequently. Such uncertainty-aware scheduling is proposed for future evaluation and is not part of the current experimental protocol.

5. Conclusion

This paper introduces a FedNova-MobileViT model for ship propeller cavitation recognition based on hydrophone spectrum sequences. A unified engineering workflow was developed to perform acoustic pre-processing, lightweight sequence representation and normalized federated optimization. Spectrogram sequences are used instead of individual frames to model the evolution of harmonic ridges and broadband collapse over time. MobileViT blocks are employed to integrate local spectral textures with wide-time-frequency dependencies. FedNova aggregation was employed to mitigate the impact of uneven local training progress among shipborne clients.

The study remains limited by its controlled five-client federation, four-state label scheme, and incomplete coverage of propeller geometries, sensor positions, seasonal sea states, hull conditions, and sensor aging. Hydrophone spectra alone also leave residual ambiguity between incipient cavitation and maneuver-induced broadband noise. The reported convergence trajectories use one initialization and sampling schedule, annotation agreement was resolved by consensus without a recorded coefficient, and embedded latency and energy values are specific to the reported hardware profile. These constraints limit fleet-wide generalization and require cautious interpretation of operational alarm thresholds.

Future work will extend the framework to multimodal federated diagnostics by aligning hydrophone spectra with rpm, thrust, vibration, and wake-condition measurements. Longer field trials should quantify transfer across vessels, seasons, hull fouling states, and sensor degradation, with repeated-seed uncertainty and vessel-held-out evaluation. Adaptive client selection will also be investigated so that rare cavitation states and persistently uncertain domains receive additional communication opportunities without routine exposure of raw acoustic recordings.

6. Declarations

6.1 Author Contributions

Leonardo Bruno contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision, project administration, and funding acquisition. Alessandro Moretti contributes to validation, analysis, investigation, data collection, draft preparation. All authors have read and agreed with the manuscript before its submission and publication.

6.2 Funding and Acknowledgments

This research received no specific financial support from any funding agency.

6.3 Institutional Review Board Statement

Not applicable.

6.4 Informed Consent Statement

Not applicable.

6.5 Data Availability

Not applicable.

6.6 Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

6.7 Declaration of generative AI and AI-assisted technologies

During the preparation of this manuscript, the authors utilized Grammarly for linguistic polishing, logical organization and draft revision. The generative AI tool was not involved in research design, data collection, data analysis, or the equation of core conclusions. All outputs generated by AI were reviewed, revised and verified manually by the authors. The authors take full responsibility for all content of this manuscript.

7. References

[1] Bocanegra, J. A., Borelli, D., Gaggero, T., Picó, R., & Tani, G. (2025). Acoustic characterization of a cavitation tunnel for ship propeller noise studies. Journal of Ocean Engineering and Science, 10(3), 330–341. https://doi.org/10.1016/j.joes.2024.02.001

[2] Tomy, J. P., Gosda, R., Berger, S., Bingham, H. B., Andersen, P., & Abdel-Maksoud, M. (2024). Modelling the acoustic radiation of propeller tip vortex cavitation in potential flow simulations. Ocean Engineering, 311, 118732. https://doi.org/10.1016/j.oceaneng.2024.118732

[3] Tan, Q., Ding, Y., Sui, C., & Xiang, L. (2024). Effects of propeller cavitation on ship propulsion performance in off-design operating conditions. Ocean Engineering, 313, 119430. https://doi.org/10.1016/j.oceaneng.2024.119430

[4] Alizadeh, V., Abbaspour, M., Mohseni Arasteh, A., Lari, K., & Torabi Azad, M. (2024). Numerical solution and experimental investigation for hydro-acoustic analysis and noise reduction assessment of ship ducted propeller. Heliyon, 10(8), e29632. https://doi.org/10.1016/j.heliyon.2024.e29632

[5] Chen, L., Luo, X., & Zhou, H. (2024). A ship-radiated noise classification method based on domain knowledge embedding and attention mechanism. Engineering Applications of Artificial Intelligence, 127, 107320. https://doi.org/10.1016/j.engappai.2023.107320

[6] Yang, S., Jin, A., Zeng, X., Wang, H., Hong, X., & Lei, M. (2024). Underwater acoustic target recognition based on sub-band concatenated Mel spectrogram and multidomain attention mechanism. Engineering Applications of Artificial Intelligence, 133, 107983. https://doi.org/10.1016/j.engappai.2024.107983

[7] Müller, N., Reermann, J., & Meisen, T. (2024). Navigating the Depths: A Comprehensive Survey of Deep Learning for Passive Underwater Acoustic Target Recognition. IEEE Access, 12, 154092–154118. https://doi.org/10.1109/access.2024.3480788

[8] Zhao, Y., Jiang, A., Jiang, W., Yang, X., Xia, X., & Gu, X. (2024). Intelligent Fault Diagnosis Method for Constant Pressure Variable Pump Based on Mel-MobileViT Lightweight Network. Journal of Marine Science and Engineering, 12(9), 1677. https://doi.org/10.3390/jmse12091677

[9] Li, J., & Yang, H. (2024). Deep learning method with auditory passive attention for underwater acoustic target recognition under the condition of ship interference. Ocean Engineering, 302, 117674. https://doi.org/10.1016/j.oceaneng.2024.117674

[10] Vyas, A., Lin, P.-C., Hwang, R.-H., & Tripathi, M. (2024). Privacy-Preserving Federated Learning for Intrusion Detection in IoT Environments: A Survey. IEEE Access, 12, 127018–127050. https://doi.org/10.1109/access.2024.3454211

[11] Ji, S., Tan, Y., Saravirta, T., Yang, Z., Liu, Y., Vasankari, L., Pan, S., Long, G., & Walid, A. (2024). Emerging trends in federated learning: from model fusion to federated X learning. International Journal of Machine Learning and Cybernetics, 15(9), 3769–3790. https://doi.org/10.1007/s13042-024-02119-1

[12] Smith, T. A., Grech La Rosa, A., Piggott, G., Gaivota, J. A. N., & McMorran, S. S. (2025). Propeller cavitation on small craft: Underwater noise measurements and visualisation from full-scale trials. Ocean Engineering, 317, 120024. https://doi.org/10.1016/j.oceaneng.2024.120024

[13] Li, G., Bu, W., & Yang, H. (2024). Noise reduction method for ship radiated noise signal based on modified uniform phase empirical mode decomposition. Measurement, 227, 114193. https://doi.org/10.1016/j.measurement.2024.114193

[14] Yang, H., Wang, C., & Li, G. (2024). Adaptive classification system of ship-radiated noise based on hybrid multi-algorithm. Ocean Engineering, 310, 118633. https://doi.org/10.1016/j.oceaneng.2024.118633

[15] Tang, J., Gao, W., Ma, E., Sun, X., & Ma, J. (2024). Deep Learning Based Underwater Acoustic Target Recognition: Introduce a Recent Temporal 2D Modeling Method. Sensors, 24(5), 1633. https://doi.org/10.3390/s24051633

[16] Khan, R., Saeed, U., & Koo, I. (2024). FedLSTM: A Federated Learning Framework for Sensor Fault Detection in Wireless Sensor Networks. Electronics, 13(24), 4907. https://doi.org/10.3390/electronics13244907

[17] Lu, Z., Pan, H., Dai, Y., Si, X., & Zhang, Y. (2024). Federated Learning With Non-IID Data: A Survey. IEEE Internet of Things Journal, 11(11), 19188–19209. https://doi.org/10.1109/jiot.2024.3376548

[18] Bang, J., Woo, S., & Lee, J. (2024). DRL-empowered joint batch size and weighted aggregation adjustment mechanism for federated learning on non-IID data. ICT Express, 10(4), 863–870. https://doi.org/10.1016/j.icte.2024.04.011

[19] Wang, K., Yu, W., & Yamauchi, T. (2024). MVT-CEAM: a lightweight MobileViT with channel expansion and attention mechanism for facial expression recognition. Signal, Image and Video Processing, 18(10), 6853–6865. https://doi.org/10.1007/s11760-024-03356-1

[20] Syu, J.-H., & Lin, J. C.-W. (2024). Heterogeneous Federated Learning for Non-IID Smartwatch Data Classification. IEEE Internet of Things Journal, 11(18), 29811–29818. https://doi.org/10.1109/jiot.2024.3410393

[21] Luo, G., Chen, N., He, J., Jin, B., Zhang, Z., & Li, Y. (2024). Privacy-preserving clustering federated learning for non-IID data. Future Generation Computer Systems, 154, 384–395. https://doi.org/10.1016/j.future.2024.01.005

[22] Zhao, X., & Shen, D. (2024). FedSW: Federated learning with adaptive sample weights. Information Sciences, 654, 119873. https://doi.org/10.1016/j.ins.2023.119873

[23] Singh, N., & Adhikari, M. (2025). SelfFed: Self-adaptive Federated Learning with Non-IID data on Heterogeneous Edge Devices for Bias Mitigation and Enhance Training Efficiency. Information Fusion, 118, 102932. https://doi.org/10.1016/j.inffus.2025.102932

[24] Lee, T., Ahn, B.-K., Lee, K., Lee, Y., & Kim, H.-J. (2024). Prediction of cavity inception speed and underwater radiated noise of a full-scale marine propeller based on a cavitation tunnel model test. Ocean Engineering, 309, 118456. https://doi.org/10.1016/j.oceaneng.2024.118456

[25] Yi, L., Wang, G., Wang, X., & Liu, X. (2024). QSFL: Two-Level Communication-Efficient Federated Learning on Mobile Edge Devices. IEEE Transactions on Services Computing, 17(6), 4166–4182. https://doi.org/10.1109/tsc.2024.3455098

[26] Gupta, C., & Khullar, V. (2025). Modality independent federated multimodal classification system detached EEG, audio and text data for IID and non-IID conditions. Biomedical Signal Processing and Control, 108, 107938. https://doi.org/10.1016/j.bspc.2025.107938

[27] Trotta, A., Montori, F., Ciabattini, L., Billi, G., Bononi, L., & Di Felice, M. (2024). Edge human activity recognition using federated learning on constrained devices. Pervasive and Mobile Computing, 104, 101972. https://doi.org/10.1016/j.pmcj.2024.101972

[28] Jin, S.-Y., Su, Y., Ma, C.-Y., Fan, Y.-X., & Tao, Z.-Y. (2025). Hierarchical feature extraction of intrinsic modes for ship-radiated noise classification. Ocean Engineering, 326, 120878. https://doi.org/10.1016/j.oceaneng.2025.120878

[29] Tong, W., Wu, K., Wang, H., Cao, L., Huang, B., Wu, D., & Antoni, J. (2024). Adaptive Weighted Envelope Spectrum: A robust spectral quantity for passive acoustic detection of underwater propeller based on spectral coherence. Mechanical Systems and Signal Processing, 212, 111265. https://doi.org/10.1016/j.ymssp.2024.111265

[30] Sun, D., Li, Y., Liu, Z., Jia, S., & Noman, K. (2024). Physics-inspired multimodal machine learning for adaptive correlation fusion based rotating machinery fault diagnosis. Information Fusion, 108, 102394. https://doi.org/10.1016/j.inffus.2024.102394