

Uncertainty-Aware Crossmodal Fusion for Classification of Animal Behavior

Ehsan Yaghoubi^{a,*}, Florian Haselbeck^a

^a*Weihenstephan-Triesdorf University of Applied Sciences, Department of Sustainable Agricultural and Energy Systems, Freising, Am Staudengarten 1, 85354, Germany*

Abstract

Artificial intelligence offers substantial potential for acoustic monitoring of animals, from welfare assessment in precision livestock farming to wildlife conservation and ecological research, where vocalizations can indicate health, stress, and social states earlier and at lower cost than manual observation. However, recordings in these settings are obtained under uncontrolled conditions, including environmental noise, reverberation, overlapping calls, and sensors that degrade without notice. As a consequence, automated classification of animal vocalizations remains challenging, and the two dominant acoustic representations show complementary limitations: raw waveforms preserve temporal microstructure but degrade under clipping and reverberation, while log-Mel spectrograms capture harmonic organization but lose phase information and are sensitive to broadband noise. To address these challenges, we propose Uncertainty-Aware Fusion (UAF), a dual-stream framework that estimates Gaussian uncertainty for each representation and fuses them via uncertainty weighting. This mechanism assigns greater weight to the more confident representation with no reliability labels required. In a cross-species, identity-based evaluation excluding all individuals seen during training, UAF (mean pooling) achieves 59.4% accuracy / 39.7% macro F1 on the 17-class SoundWel pig vocalization benchmark and 73.1% accuracy / 71.5% macro F1 on the 3-class DogBark dataset, outperforming static-concatenation fusion by 15.7% and 20.4% relative macro F1, respectively. Ablations over four temporal aggregation strategies show that uncertainty fusion, rather than the temporal

*Corresponding author.

Email addresses: ehsan.yaghoubi@hswt.de (Ehsan Yaghoubi),
florian.haselbeck@hswt.de (Florian Haselbeck)

characteristics of animal calls, is the primary driver of the performance gain.

Keywords: Uncertainty Analysis, Multimodal Fusion, Bioacoustics, Animal Welfare, Precision Livestock Farming.

1. Introduction

Precision livestock farming employs sensor-based monitoring to assess health, stress, and welfare, often at the level of the individual animal (Berckmans, 2017). Acoustic sensing is a widely used modality, since it requires no animal contact, covers a pen from a fixed position, and adds comparatively little hardware per monitored unit (Coutant et al., 2024; Shorten and Welten, 2021; Shorten and Hunter, 2023). Vocalizations have furthermore been associated with pain, thermal discomfort, hunger, and social conflict (Manteuffel et al., 2004). In recent years, deep learning-based approaches trained on raw audio or spectrogram input have reached high accuracy on curated datasets, in welfare assessment (Coutant et al., 2024) as well as in wildlife conservation (Kershenbaum et al., 2025) and ecological research (Penar et al., 2020). This performance is largely obtained under favorable acquisition conditions and is often not reproduced in deployment, where signal quality is a limiting factor: recordings contain machinery and environmental noise, reverberation from hard surfaces, and overlapping calls from multiple animals, and sensors degrade gradually without this being detected at the time of recording. Since such predictions inform subsequent management decisions, quantifying the reliability of a prediction is of interest in addition to classification performance.

Generally, there are two dominant acoustic representations, namely raw time-domain waveforms and log-Mel frequency spectrograms. While processing waveform signals preserves temporal microstructure and phase, it is sensitive to clipping and reverberation (Baevski et al., 2020). On the other hand, spectrogram-based models capture harmonic and frequency organization but discard phase and degrade under broadband noise (Stowell, 2022). Therefore, neither representation is uniformly superior because the relative reliability shifts with recording conditions, which is a practically unavoidable reality in uncontrolled environments and field deployments. Existing dual-representation approaches (Kong et al., 2020; Liang et al., 2022) demonstrate that combining waveform and spectrogram features consistently outperforms either stream alone. However, these methods usually employ *static fusion*:

both representations receive fixed weights regardless of their reliability for a given input. Under the variable acoustic conditions of real-world animal monitoring, static weights propagate noise from the degraded representation rather than suppressing it. Although uncertainty-aware fusion has been explored in audio-visual human emotion recognition (Tellamekala et al., 2023; Xie et al., 2024b), no prior work has adapted reliability-driven weighting to audio dual-representation fusion for bioacoustic classification. This is challenging because, unlike audio-visual fusion, both modalities originate from the same acoustic signal, requiring the model to determine which representation should be trusted in each step in time.

We address this gap with **Uncertainty-Aware Fusion (UAF)**, a framework that dynamically weights waveform and spectrogram contributions using per-utterance latent uncertainty estimates (Fig. 1). First, modality-specific ResNet-18 encoders map each input to a temporal feature sequence, and a temporal aggregation module produces a single utterance-level embedding per modality. An uncertainty head then parameterizes a per-utterance diagonal Gaussian $\mathcal{N}(\boldsymbol{\mu}_m, \boldsymbol{\sigma}_m^2)$ for each modality $m \in \{A, F\}$, and dimension-wise uncertainty weights fuse the two means: each latent dimension is dominated by whichever modality is more confident for that input. Training optimizes a multi-term objective that learns from three streams, i.e., waveform, spectrogram, and a fusion branch. We evaluate on two publicly available bioacoustic benchmark datasets, i.e., SoundWel, a 17-class pig welfare vocalization (Briefer et al., 2022) and DogBark, a dataset for context-dependent dog bark classification (Yin and McCowan, 2004). In summary, our contributions are threefold:

- We propose a dual-representation architecture for bioacoustic classification that processes raw waveform and log-Mel spectrogram inputs in parallel, capturing temporal and frequency cues that neither representation provides on its own.
- We employ an uncertainty-aware fusion mechanism that estimates the reliability of each representation at every time step and weights their contributions accordingly. This directly targets a practical problem in real-world and barn recordings, where noise, clipping, or occlusion can degrade one representation while leaving the other intact.
- We evaluate our framework on two bioacoustic benchmarks, SoundWel (pig, 17 classes) and DogBark (dog, 3 classes), and report ablation

studies that show the gain from uncertainty-aware fusion relative to static concatenation baselines.

2. Related Work

2.1. Acoustic Representations in Animal Bioacoustics

The time-domain waveform and the frequency spectrogram are two different representations of acoustic signals that can be used for behavioral classification of animals. Early approaches relied on hand-crafted features for specific species and monitoring objectives (Coutant et al., 2024; McLoughlin et al., 2019). These features are grouped into three types: time-domain descriptors such as call duration and inter-call intervals; frequency-domain descriptors including fundamental frequency, formants, and spectral centroid; and energy-related measures such as amplitude and signal-to-noise ratio. Dynamic features such as jitter and shimmer capture temporal micro-variations, while derived representations such as Mel-Frequency Cepstral Coefficients (MFCCs) and their temporal derivatives have become common in automatic classification systems (Li et al., 2024). For instance, pig distress calls exhibit higher pitch, longer duration, and noisier spectra, whereas chicken alarm calls are typically short, high-pitched bursts (Coutant et al., 2024). While hand-crafted features have enabled considerable progress in livestock monitoring (McLoughlin et al., 2019) and wildlife conservation, they require substantial domain expertise, may not generalize well across recording conditions or species, and can miss subtle, nonlinear interactions between temporal and spectral information. These limitations have motivated a shift toward learned representations via deep neural networks; however, whether time-domain waveforms, frequency-domain spectrograms, or both yield the best learned features remains an open question with distinct trade-offs in information preservation, computational cost, and noise robustness.

2.2. Spectrogram-based Deep Learning Models

Spectrogram-based approaches have emerged as the dominant paradigm in bioacoustic deep learning, converting raw audio signals into time-frequency representations such as mel-spectrograms, MFCCs, or delta-mel features before applying, e.g., convolutional neural networks (CNNs) (Stowell, 2022; Xie et al., 2024a). By treating spectrograms as images, computer vision approaches can learn hierarchical spatial patterns that capture both temporal and spectral dynamics, consistently outperforming traditional classifiers based

on hand-crafted features in comparative studies (Ghani et al., 2023; Bergler et al., 2022). Hybrid architectures that incorporate recurrent units, e.g., CNN-LSTM or CNN-GRU, further enhance performance by modeling temporal dependencies across spectrogram frames (Gupta et al., 2021). More recently, transformer-based models that treat the log-mel spectrogram as a sequence of 2-D patches, such as the Audio Spectrogram Transformer (AST) (Gong et al., 2021), achieve state-of-the-art results on large-scale audio benchmarks. The AVES (Animal Vocalization Encoder based on Self-Supervision) model (Hagiwara, 2023) extends this paradigm to animal vocalizations via self-supervised transformer pre-training, yielding strong transfer performance on bioacoustic classification.

These spectrogram-based models have been successfully applied across diverse bioacoustic domains. In livestock applications, automated classification systems using multiple CNN backbones (e.g., AlexNet, ResNet, EfficientNet) have demonstrated high accuracy for dog vocalizations (Gómez-Armenta et al., 2024), while emotion recognition systems show promise for farm animal welfare monitoring (Briefer et al., 2022). For wildlife monitoring, large-scale species classification of birds and insects has leveraged CNNs to capture discriminative frequency patterns across taxa (Nanni et al., 2020; Stowell, 2022), while in marine bioacoustics, deep learning models have been developed for detecting baleen whale vocalizations (Schall et al., 2024). Recent frameworks such as ANIMAL-SPOT (Bergler et al., 2022) provide end-to-end pipelines for detection and classification directly from spectrograms, showing strong generalization across multiple species and call types.

Despite these empirical successes, spectrogram-based methods have known limitations. The time-frequency transformation inherently discards phase information, which may contain discriminative temporal microstructures relevant for certain call types. Model performance is sensitive to spectrogram resolution and preprocessing choices such as window size, hop length, and frequency binning, requiring careful tuning for each application. Furthermore, performance often degrades when deployed in real-world scenarios that differ from training conditions, particularly under varying noise profiles or recording equipment (Stowell et al., 2019; Shorten and Hunter, 2023). These robustness challenges suggest that relying solely on frequency-domain representations could be insufficient for generalizable bioacoustic monitoring systems.

2.3. Raw Waveform and Hybrid Audio Models

Motivated by the potential information loss in spectrogram conversion, raw waveform modeling has gained attention as a complementary representation strategy. Raw waveform approaches often apply convolutional filters directly to time-domain audio signals, allowing the network to learn discriminative temporal features without explicit frequency transformation (Tokozume and Harada, 2017). Self-supervised pre-training on raw waveforms—exemplified by wav2vec 2.0 (Baevski et al., 2020) and HuBERT (Hsu et al., 2021)—has demonstrated that contextual representations learned from large unlabeled audio corpora transfer effectively to downstream classification tasks. Large-scale pretrained models such as PANNs (Pretrained Audio Neural Networks) further show that fusing waveform- and spectrogram-level features improves audio pattern recognition. The Wavegram-Logmel-CNN variant in PANNs concatenates features from both domains and achieves better performance than either branch alone (Kong et al., 2020). The theoretical advantage of waveform models lies in preserving complete signal information, including phase relationships and fine-grained temporal structures that may be smoothed or lost during spectrogram computation.

In bioacoustics, however, raw waveform modeling remains relatively underexplored compared to spectrogram-based approaches. Bravo Sanchez et al. demonstrate that SincNet-based waveform architectures can classify avian calls without spectrograms, establishing the viability of end-to-end waveform learning for bioacoustic tasks (Bravo Sanchez et al., 2021). Nevertheless, bioacoustic applications continue to face unique challenges, including data scarcity, high intra-species variability, and diverse recording conditions. More critically, although both representations offer complementary information, existing multi-representation approaches such as PANNs (Kong et al., 2020) as well as the dual-stream waveform and log-mel fusion of Liang et al. (2022) rely on static, fixed-weight concatenation that does not adapt when one representation becomes less reliable under specific conditions. The key unresolved challenge is not whether to combine both representations but how to dynamically assess and respond to their relative reliability.

2.4. Multimodal Fusion Strategies

The recognition that single representations may be insufficient has led to growing interest in multimodal integration, both across different sensor types and representations of the same audio signal. Baltrušaitis et al. provide the foundational taxonomy of multimodal learning strategies, distinguishing early

fusion, late fusion, and intermediate fusion, and highlighting representation learning and fusion as the two central challenges (Baltrušaitis et al., 2019). In livestock monitoring, recent systems have combined acoustic signals with accelerometry and inertial sensors to improve the classification of behavioral states and welfare indicators (Vu et al., 2024; Ferrero et al., 2025). These multimodal systems leverage complementary information from different sensing modalities to provide more robust predictions in noisy farm environments, where visual occlusion, acoustic interference, or sensor failures can degrade individual channels.

However, most existing fusion strategies employ simple integration methods that assume equal reliability across modalities. Early fusion approaches concatenate raw features or low-level representations before classification, while late fusion combines independent predictions from modality-specific classifiers through averaging or voting. Intermediate strategies extract separate embeddings from each modality and concatenate them before the final classification layers (Baltrušaitis et al., 2019). These methods treat all modalities equally informative at all times, assigning fixed or learned but static weights that do not adapt to temporal variations in signal quality. Peng et al. show empirically that joint multimodal training suffers from a modality-imbalance problem and propose on-the-fly gradient modulation as a solution (Peng et al., 2022).

This assumption of uniform reliability is problematic in real-world deployments, where waveform and spectrogram representations show input-dependent failure modes. Existing fusion methods amplify the noise from a degraded representation rather than suppress it, motivating reliability-adaptive weighting.

2.5. Uncertainty and Reliability Modeling in Multimodal Learning

The concept of uncertainty-aware fusion, where modality contributions are dynamically weighted based on estimated reliability, has roots in Bayesian deep learning. Gal and Ghahramani showed that Monte Carlo dropout provides a computationally tractable approximation to Bayesian inference, yielding epistemic uncertainty estimates at test time without modifying the training procedure (Gal and Ghahramani, 2016), while Kendall and Gal formalized the canonical decomposition into aleatoric uncertainty (irreducible data noise) and epistemic uncertainty (model uncertainty reducible with more data) (Kendall and Gal, 2017). In the multimodal domain, Subedar et al. (2019) applied deep Bayesian variational inference to audio-visual activity recognition and

demonstrated that per-modality uncertainty estimates improve robustness when one sensor stream is corrupted. Han et al. (2023) extended uncertainty-driven fusion to arbitrary numbers of views via Dempster-Shafer evidence fusion, enabling reliable classification under missing or degraded inputs. In multimodal emotion recognition, the COLD Fusion framework (Tellamekala et al., 2023) introduced a principled approach by learning latent distributions over unimodal temporal contexts and constraining their variance to represent how informative each modality is toward the prediction task. By imposing calibration and ordinal ranking constraints on modality-wise uncertainty estimates, COLD Fusion enables adaptive weighting that favors more reliable modalities at each time step.

3. Methods

The core idea of the proposed Uncertainty-Aware Fusion (UAF) framework is to replace static fusion weights for time-domain waveform and spectrogram representations with adaptive weights estimated directly from the learned features. This enables the model to quantify modality-specific uncertainty and suppress less reliable representations without requiring explicit reliability annotations. We first formulate the classification task and establish the notation used throughout the paper. We then describe the proposed framework in detail, including the overall architecture, unimodal waveform and spectrogram encoders, temporal aggregation, uncertainty estimation, the adaptive fusion strategy, and the corresponding training objective.

3.1. Problem Definition

The task we address in this paper is to classify animal vocalizations into K discrete behavioral categories. Given an audio recording, two complementary representations are extracted: a raw waveform $x_A \in \mathbb{R}^L$ and a three-channel stacked spectrogram $x_F \in \mathbb{R}^{C \times H \times W}$, where L is the waveform length, $C = 3$ channels encode log-Mel, Δ -Mel, and $\Delta\Delta$ -Mel coefficients, and $H \times W$ is the spectrogram spatial resolution (Young et al., 2002; Takahashi et al., 2016). We consider three-channel spectrogram input rather than the one-channel log-Mel because the derivative channels capture short-term spectral dynamics (onset, offset, and formant-transition cues) that the log-Mel spectrogram discards (Young et al., 2002). The goal is to learn a mapping $f : (x_A, x_F) \mapsto \hat{y} \in \{1, \dots, K\}$ that minimizes classification error while producing calibrated confidence estimates.

3.2. Architecture Overview

Figure 1 illustrates the proposed Uncertainty-Aware Fusion (UAF) framework, which we realize in four steps. First, we perform **modality encoding** by passing each modality through a dedicated ResNet-18 encoder, producing a temporal sequence of T feature vectors of dimension $C_e = 512$. These features are compressed by a temporal aggregation module into a single utterance-level embedding $\mathbf{h}_m \in \mathbb{R}^{C_e}$ (see Sections 3.3 and 3.4). Then, we perform **uncertainty parameterization** by mapping each embedding to the mean and variance of a diagonal Gaussian, $q_m(z | \mathbf{h}_m) = \mathcal{N}(\boldsymbol{\mu}_m, \text{diag}(\boldsymbol{\sigma}_m^2))$ for $m \in A, F$, where A and F denote the waveform and spectrogram modalities. We use the predicted variance as an implicit measure of modality reliability, enabling the model to adaptively weight each modality without requiring explicit reliability annotations (see Section 3.5). Next, in **uncertainty-weighted fusion** step, we combine these Gaussians dimension-wise, giving more weight to whichever modality has lower variance in each latent dimension, and obtain a fused latent sample $\mathbf{z} \in \mathbb{R}^{D_z}$ (see Section 3.6). Finally, we do a **joint classification** by applying a linear classifier to fused representation $\mathbf{z}$ to obtain the final prediction, while two auxiliary classifiers are applied to $\boldsymbol{\mu}_A$ and $\boldsymbol{\mu}_F$ to provide modality-specific supervision. We train the whole model with a loss function inspired by (Tellamekala et al., 2023) and (Xie et al., 2024b) that couples the predicted variances with the observed classification difficulty (see Section 3.7).

3.3. Dual Audio Encoder

We use separate encoders for the raw waveform and the spectrogram:

Waveform encoder. We use a custom 1-D ResNet-18 to map a single-channel waveform segment of length L (the number of raw audio samples) to a temporal feature map:

$$\mathbf{E}_A : \mathbb{R}^L \rightarrow \mathbb{R}^{C_e \times T}, \quad T = \lfloor L/32 \rfloor, \quad (1)$$

where C_e is the number of feature channels the encoder produces at each time step and T is the resulting sequence length (the number of temporal feature vectors handed to the aggregation module, see Section 3.4). The 32-fold downsampling from L to T accumulates over a stride-2 stem (conv 7×1 + max-pool) followed by three stride-2 residual stages. The values of C_e , L , or T are not intrinsic to the fusion mechanism: $C_e = 512$ reflects the standard

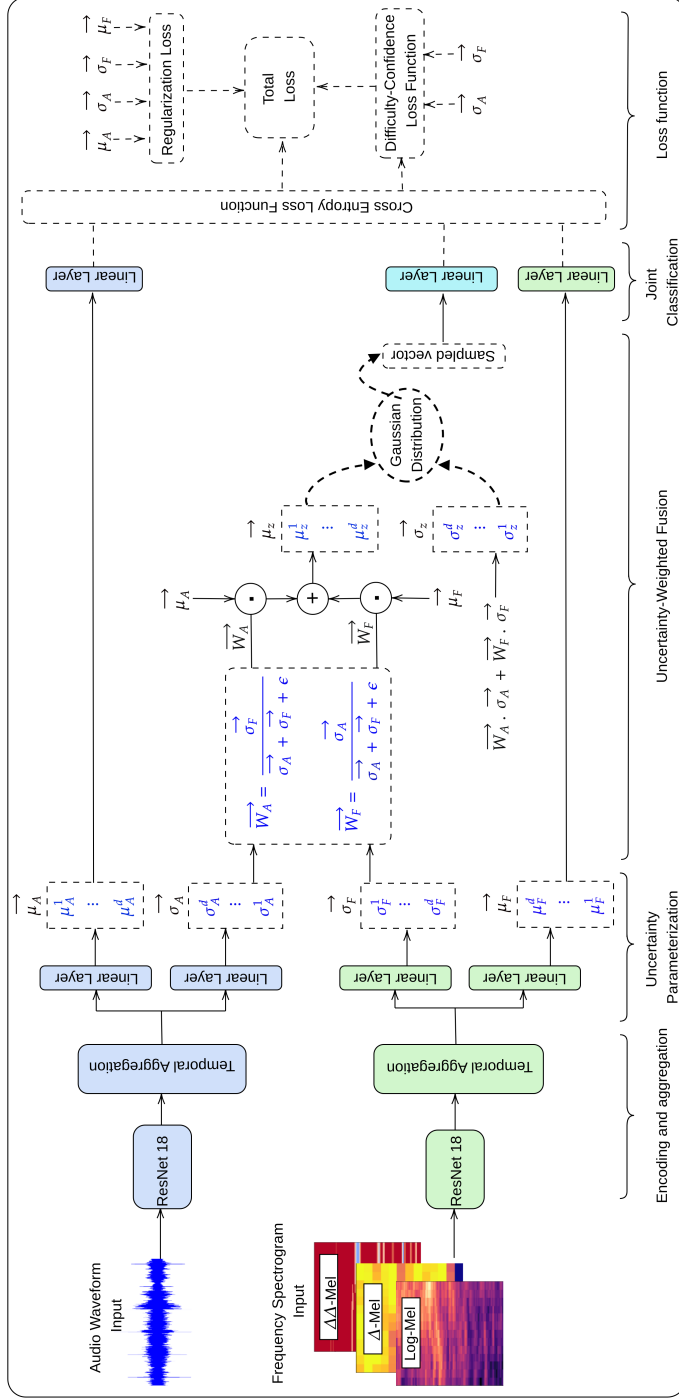


Figure 1: Overview of the proposed Uncertainty-Aware Fusion (UAF) framework for animal vocalization classification. Audio waveform and frequency spectrogram are encoded separately and temporally aggregated. The resulting embeddings are mapped to Gaussian means and variances, and the uncertainty-aware fusion module combines the two modalities into a single fused distribution by assigning higher weights to the more certain modality. A vector sampled from this fused distribution, together with two auxiliary unimodal predictions, is used for the final classification. A difficulty-confidence loss couples each modality’s uncertainty to how hard its predictions are, while a regularization term keeps the uncertainties feasible for an end-to-end train.

ResNet-18 backbone, $L = 6,400$ results from sampling a raw audio waveform at 16 kHz, and the sampling rate (400 ms) was chosen to match the typical duration of a single vocalization event in our target datasets. Employing a different backbone, segment length, or sampling rate would rescale these values without altering the underlying uncertainty-aware fusion method.

Spectrogram encoder. We use a standard ResNet-18 (He et al., 2016) to produce frequency features: $\mathbf{E}_F : \mathbb{R}^{C \times H \times W} \rightarrow \mathbb{R}^{C_e \times T}$. This means that the backbone maps $(B, C, H, W) \rightarrow (B, C_e, H/32, W/32)$, then an average pooling collapses the frequency axis, and linear interpolation aligns the temporal axis to T . With $H = W = 128$ the backbone yields a 4×4 spatial map after frequency pooling. There are four distinct temporal observations, which are linearly interpolated to $T = 200$ to match the waveform encoder’s output shape (B, C_e, T) .

3.4. Temporal Context Aggregation

Each encoder sequence $\{e_{m,t}\}_{t=1}^T \subset \mathbb{R}^{C_e}$ is aggregated into a single utterance-level embedding $\mathbf{h}_m \in \mathbb{R}^{D_h}$. To yield this embedding, our proposed UAF model applies global average pooling:

$$\mathbf{h}_m = \frac{1}{T} \sum_{t=1}^T e_{m,t}, \quad D_h = C_e = 512. \quad (2)$$

We additionally evaluate three alternative aggregation modules as ablations of this design choice:

- *GRU*: A multi-layer GRU processes the sequence in temporal order, and $\mathbf{h}_m$ is taken as the final hidden state of the last recurrent layer.
- *Transformer*: A Pre-Layer Normalization Transformer encoder with learnable positional embeddings operates on the time step sequence, and the output is mean-pooled to $\mathbf{h}_m \in \mathbb{R}^{C_e}$.
- *Attention pooling*: A two-layer MLP processes the embeddings of each time step, and a softmax over the sequence turns the output into attention weights such that the time steps that are more informative contribute more.

For all variants, $\mathbf{h}_m$ is further processed by the downstream uncertainty head and fusion module, isolating temporal aggregation as the critical architectural design in the ablation.

3.5. Gaussian Uncertainty Parameterization

The gaussian uncertainty parameterization is the major part where our approach differs from a conventional dual-stream classifier. Instead of treating each modality’s utterance embedding $\mathbf{h}_m$ as a fixed point, an uncertainty head maps it to a distribution, producing a mean $\boldsymbol{\mu}_m$ and a variance $\boldsymbol{\sigma}_m^2$. The mean plays the role of a conventional embedding, while the variance indicates the model’s confidence in that estimate. Therefore, the variance should be small along dimensions where a modality is clear and unambiguous, and it should be large along dimensions where the input is noisy or degraded. As there are no reliability labels, the association between uncertainty and input quality is learned using a loss function inspired by (Tellamekala et al., 2023) and (Xie et al., 2024b). We explain the loss function and its empirical evaluation in Sections 3.7 and 5.3, respectively.

We use two independent linear layers to yield $\boldsymbol{\mu}_m$ and $\boldsymbol{\sigma}_m$ from the embeddings of the above-described waveform and spectrogram encoders:

$$\boldsymbol{\mu}_m = W_\mu^{(m)} \mathbf{h}_m, \quad \boldsymbol{\sigma}_m = \text{softplus}(W_\sigma^{(m)} \mathbf{h}_m) + \varepsilon, \quad (3)$$

where $W_\mu^{(m)}, W_\sigma^{(m)} \in \mathbb{R}^{D_z \times D_h}$ are learnable projections. Softplus keeps $\boldsymbol{\sigma}_m$ positive while remaining differentiable everywhere, such that it can be trained end-to-end with standard backpropagation. Adding the small constant $\varepsilon = 10^{-6}$ guarantees $\boldsymbol{\sigma}_m$ never reaches exactly zero, which protects the fusion rule given in Section 3.6 from dividing by zero. We consider the dimensionality of $\boldsymbol{\mu}_m$ and $\boldsymbol{\sigma}_m$ as a hyperparameter (D_z) and optimize it using Bayesian optimization leveraging the Python framework Optuna (Akiba et al., 2019) (see Section 4.3).

3.6. Dimension-wise Uncertainty Fusion

Given the Gaussian parameters $(\boldsymbol{\mu}_A, \boldsymbol{\sigma}_A)$ and $(\boldsymbol{\mu}_F, \boldsymbol{\sigma}_F)$, fusion weights are computed independently for each latent dimension $d \in \{1, \dots, D_z\}$:

$$w_A(d) = \frac{\sigma_F(d)}{\sigma_A(d) + \sigma_F(d)}, \quad w_F(d) = \frac{\sigma_A(d)}{\sigma_A(d) + \sigma_F(d)}. \quad (4)$$

Consequently, a dimension in which the spectrogram is highly uncertain ($\sigma_F(d) \gg \sigma_A(d)$) receives proportionally more weight from the waveform, and vice versa. Stacking these per-dimension weights into vectors $\mathbf{w}_A, \mathbf{w}_F \in \mathbb{R}^{D_z}$

(with $\mathbf{w}_A + \mathbf{w}_F = \mathbf{1}$), the fused Gaussian parameters are defined as:

$$\boldsymbol{\mu}_z = \mathbf{w}_A \odot \boldsymbol{\mu}_A + \mathbf{w}_F \odot \boldsymbol{\mu}_F, \quad (5)$$

$$\boldsymbol{\sigma}_z = \mathbf{w}_A \odot \boldsymbol{\sigma}_A + \mathbf{w}_F \odot \boldsymbol{\sigma}_F. \quad (6)$$

During training, a latent sample is drawn via the reparameterization trick:

$$\mathbf{z} = \boldsymbol{\mu}_z + \boldsymbol{\sigma}_z \odot \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \quad (7)$$

At inference, we set $\mathbf{z} = \boldsymbol{\mu}_z$. The vector $\mathbf{z}$ is then further processed by a dropout-regularized linear layer to produce class logits $\hat{\ell}_z = W_c \mathbf{z} \in \mathbb{R}^K$.

3.7. Training Objective

Inspired by Tellamekala et al. (2023) and Xie et al. (2024b), we design a novel loss function that supervises classification from the fused and unimodal predictions, and simultaneously learns the uncertainty to help a dynamic fusion based on the confidence of the modalities. In the following, we explain each component of this loss function.

Classification loss. Each unimodal stream shown in Fig. 1 has an independent linear head on top of $\boldsymbol{\mu}_A$ and $\boldsymbol{\mu}_F$, allowing it to learn the class label directly from that modality. The fused branch (logits from $\mathbf{z}$) is additionally supervised using the same class label. Together, the three branches form the classification loss:

$$\mathcal{L}_{\text{cls}} = \frac{1}{3} \left[\mathcal{H}(\hat{\ell}_z, y) + \mathcal{H}(\hat{\ell}_A, y) + \mathcal{H}(\hat{\ell}_F, y) \right], \quad (8)$$

where $\mathcal{H}(\cdot, y)$ denotes cross-entropy with the ground-truth label y , and $\hat{\ell}_m = W_m \boldsymbol{\mu}_m$. During training, $\hat{\ell}_z$ is computed from the reparameterized sample, see Eq. 7, while in the inference phase, the unimodal classifiers receive means $\boldsymbol{\mu}_A$ and $\boldsymbol{\mu}_F$ without the reparameterization noise.

Difficulty-Confidence loss. The fusion process requires uncertainty scores. However, no ground-truth labels for uncertainty are available. Therefore, we compare within each mini-batch of B samples how hard a sample actually was to classify against how confident the model was about its prediction. This allows the model to learn meaningful uncertainty estimates from relative comparisons within a batch.

For each modality $m \in \{A, F\}$ and sample i , we compute a difficulty score (the sample’s own cross-entropy loss) and a confidence score (the inverse of its predicted uncertainty):

$$d_m^{(i)} = \mathcal{H}(\hat{\ell}_m^{(i)}, y^{(i)}), \quad s_m^{(i)} = \frac{1}{\|\boldsymbol{\sigma}_m^{(i)}\|_2 + \varepsilon}. \quad (9)$$

Both scores are then normalized via a softmax with temperature τ over all B samples into two probability distributions describing how difficult and how confident each sample is *relative to the rest of the batch*:

$$\mathbf{p}_D^m = \text{softmax}(\mathbf{d}_m / \tau), \quad \mathbf{p}_S^m = \text{softmax}(\mathbf{s}_m / \tau). \quad (10)$$

Matching these two distributions via a symmetric KL divergence encourages samples that rank as relatively difficult under $\mathbf{p}_D^m$ to also rank as relatively confident under $\mathbf{p}_S^m$, which corresponds to lower $\boldsymbol{\sigma}_m$. In contrast, easy samples are encouraged to have lower confidence (higher $\boldsymbol{\sigma}_m$). Thus, the loss couples relative difficulty to relative confidence. The results in Section 5.3 show the correlation between $\|\sigma_z\|$ and the misclassification error. For each modality m , this correlation is imposed with the symmetric KL divergence:

$$\mathcal{L}_{\text{DC},m} = \text{KL}(\mathbf{p}_D^m \parallel \mathbf{p}_S^m) + \text{KL}(\mathbf{p}_S^m \parallel \mathbf{p}_D^m), \quad m \in \{A, F\}. \quad (11)$$

To apply the same idea jointly (as a single loss term) across both modalities, we interleave the modalities’ difficulty and confidence scores into $2B$ -length vectors:

$$\mathbf{d}_{AF} = [d_A^{(1)}, d_F^{(1)}, \dots, d_A^{(B)}, d_F^{(B)}], \quad \mathbf{s}_{AF} = [s_A^{(1)}, s_F^{(1)}, \dots, s_A^{(B)}, s_F^{(B)}]. \quad (12)$$

The same symmetric KL penalty is then applied to this combined set of $2B$ samples:

$$\mathcal{L}_{\text{DC},AF} = \text{KL}(\mathbf{p}_D^{AF} \parallel \mathbf{p}_S^{AF}) + \text{KL}(\mathbf{p}_S^{AF} \parallel \mathbf{p}_D^{AF}). \quad (13)$$

Variance regularizer. To prevent variance collapse (the model pushing $\boldsymbol{\sigma}_m \rightarrow 0$ to suppress noisy gradients), we penalize departure from an isotropic standard normal prior (Tellamekala et al., 2023; Xie et al., 2024b):

$$\begin{aligned} \mathcal{L}_R &= \sum_{m \in \{A, F\}} \text{KL}(\mathcal{N}(\boldsymbol{\mu}_m, \boldsymbol{\sigma}_m^2) \parallel \mathcal{N}(\mathbf{0}, \mathbf{I})) \\ &= -\frac{1}{2} \sum_m \sum_{d=1}^{D_z} \left[1 + \log \sigma_{m,d}^2 - \mu_{m,d}^2 - \sigma_{m,d}^2 \right], \end{aligned} \quad (14)$$

averaged over the batch.

The total training objective composed of the above-described components is finally formulated as:

$$\mathcal{L} = \lambda_{\text{cls}} \mathcal{L}_{\text{cls}} + \lambda_{\text{DC},A} \mathcal{L}_{\text{DC},A} + \lambda_{\text{DC},F} \mathcal{L}_{\text{DC},F} + \lambda_{\text{DC},AF} \mathcal{L}_{\text{DC},AF} + \lambda_R \mathcal{L}_R, \quad (15)$$

where default weights are set empirically to $\lambda_{\text{cls}} = 1$, $\lambda_{\text{DC},A} = \lambda_{\text{DC},F} = \lambda_{\text{DC},AF} = 10^{-3}$, $\lambda_R = 10^{-4}$.

4. Experimental Setup

4.1. Datasets and Data Splits

We evaluate the proposed uncertainty-aware fusion framework on two publicly available datasets representing (i) different animal vocalization classification tasks to assess cross-species generalization and (ii) different task complexities: pig calls and dog barks context classification.

SoundWel Dataset. The SoundWel dataset (Briefer et al., 2022) is a multi-institutional collection of pig vocalizations under the SoundWel research program. It contains 6,887 labeled audio clips recorded by six independent research teams across Europe (ETHZ, FBN, IASPA, IASPB, IASPC, NMBU), covering three age groups (piglet, weaner, growing/finishing) and 17 behavioral context categories spanning welfare-negative states (e.g. isolation, castration, crushing), welfare-positive states (e.g. enriched, reunion), and nursing-related behaviors. The dataset is severely class-imbalanced: For instance, isolation accounts for 29.7% of samples, while surprise accounts for only 0.2%. Individual clip durations have a mean of 0.36 s and a maximum of 3.6 s.

DogBark Dataset. The DogBark dataset (Yin and McCowan, 2004) contains dog vocalizations from ten individual dogs spanning six breeds, annotated across three behavioral contexts: aggression (14.3%), contact (55.6%), and play (30.2%). The original dataset comprises 4,672 samples; however, only 693 labeled recordings are publicly available.

Data Splits. A random stratified split would cause the recordings of the same individual to appear in both training and test sets, encouraging the model to exploit individual-specific vocal fingerprints rather than context-discriminative features. This is known as data leakage between train and test sets, which

unrealistically inflates the model’s performance during the test phase (Bennett et al., 2024). Instead, in this work, we considered identity-aware splits for both datasets. However, results under a random stratified split are also provided in Table B.1 (Appendix B) for comparison with the previously published result of Briefer et al. (2022).

The **SoundWel** dataset does not provide an identity train-val-test split, and only a limited number of samples have a pig ID. To avoid data leakage, we used the available metadata, i.e., filename patterns encoding per-animal or per-group identifiers (pen IDs, animal IDs, sow IDs, or litter numbers depending on team), recording year to disambiguate cohorts across experimental sessions, and recording team as a coarse grouping variable to prevent cross-team leakage. Recordings for which no parseable identity could be inferred (2,398 recordings, 34.8% of the dataset) are assigned to the training set such that the model could benefit from learning on more samples without the risk of data leakage. The resulting split is: train 5,452 (79.2%) | val 677 (9.8%) | test 758 (11.0%). All six recording teams contribute to every subset, and 13 of the 17 classes appear in the test set. The four absent classes in the test set (BeforeNursing, Fighting, Huddling, Run) have no recordings with parseable pig identities. Therefore, the macro-averaged F1-score is computed over the 13 classes present in the test partition.

For the **DogBark** dataset, we ensure that individual dogs are assigned exclusively to one partition. The test set consists of Mac (German shorthaired pointer, male) and Zoe (Australian cattle dog, female), providing cross-breed and cross-sex generalization with all three behavioral classes represented. The validation set consists of Keri and Louie; the remaining six dogs form the training set. This data split leads to 458 train (66.1%), 75 val (10.8%), and 160 test (23.1%) samples. The class distribution is nearly identical across partitions, making the macro-averaged F1-score a fair evaluation metric.

4.2. Preprocessing and Data Augmentation

Audio Preprocessing. All audio recordings are resampled to 16 kHz sampling rate, which captures the relevant acoustic bandwidth for both pig and dog vocalizations (predominantly below 8 kHz) while reducing computational cost compared to higher sampling rates. This sampling rate aligns with standard practices in animal bioacoustics and enables stable training (Stowell, 2022).

For the waveform input (x_A), raw audio signals are peak-normalized by dividing by the absolute maximum amplitude, mapping each clip to a range of $[-1, 1]$. For the spectrogram input (x_F), we compute a three-channel

time-frequency representation using a Mel-filterbank applied to the power spectrogram (STFT with FFT size $n_{\text{fft}} = 1024$, hop length = 256 samples (16 ms), Hann window, and power = 2). We extract 128 Mel-frequency bands spanning 0–8 kHz and convert to decibels (dB) to obtain the log-power Mel spectrogram. We then compute its first-order (Δ -Mel) and second-order temporal derivative ($\Delta\Delta$ -Mel). These three representations are stacked as input channels and resized to 128×128 pixels. Each channel is independently min-max normalized to $[0, 1]$.

Temporal Segmentation. As illustrated in Fig. 2, we use different temporal segmentation strategies during training and testing. During training, each audio recording is randomly cropped to a single 400 ms segment, which serves as an independent training sample. Recordings shorter than 400 ms are zero-padded: during training, the audio is placed at a random position within the target window; during validation and test, it is center-aligned. This random cropping acts as a form of data augmentation, exposing the model to different temporal regions of the same vocalization across training epochs and improving robustness to temporal variability.

The choice of 400 ms segments provides enough temporal context for both datasets and is motivated by bioacoustic considerations: Individual pig vocalizations have typical durations ranging from 0.3 to 1.8 seconds (Briefer et al., 2022). Hence, a 400 ms window likely captures the most discriminative part of a call without padding artifacts for the majority of samples. For DogBark, individual bark events are also short (typically well under one second), and the 400 ms window likely captures a complete bark unit. Shorter segments risk capturing only partial vocal units, reducing temporal context, while segments longer than a single call risk mixing adjacent events or silent gaps.

Data Augmentation. To improve model robustness to acoustic variability in realistic farm environments and prevent overfitting, we apply additive Gaussian noise during training. For each training sample, noise is generated from a standard normal distribution and scaled to achieve a signal-to-noise ratio (SNR) uniformly sampled from the range $[-5, 15]$ dB.

4.3. Training Procedure

The training process consists of two stages: hyperparameter optimization followed by a final training.

Stage 1: Hyperparameter Optimization. We use Bayesian optimization implemented via the Python framework Optuna (Akiba et al., 2019) with a Tree-structured Parzen Estimator (TPE) sampler and a median pruner to run 50 trials. Each trial trains and evaluates the model using the previously described training and validation data. To prevent further class imbalance between folds for the SoundWel dataset, we did not employ cross-validation. For DogBark, we adopted the same strategy for consistency. Training runs for up to 300 epochs with early stopping (patience 20, monitored on validation macro-averaged F1-scores). The trial that yields the highest score is selected for final training. The hyperparameter search space covers: learning rate $\in [10^{-5}, 10^{-3}]$, latent dimension $D \in \{64, 128, 256\}$, softmax temperature $\tau \in [0.5, 2.0]$, batch size $\in \{16, 32, 64\}$, weight decay $\in [10^{-6}, 10^{-4}]$, classification loss weight $\lambda_{\text{cls}} \in [0.5, 2.0]$, intra-modal DC loss weights $\lambda_{\text{DC_a}}, \lambda_{\text{DC_f}} \in [2 \times 10^{-3}, 2 \times 10^{-1}]$, cross-modal DC loss weight $\lambda_{\text{DC_af}} \in [2 \times 10^{-3}, 2 \times 10^{-1}]$, and variance regularization weight $\lambda_{\text{regu}} \in [2 \times 10^{-5}, 2 \times 10^{-3}]$.

Stage 2: Final Model Training. Using the best hyperparameter configuration from Stage 1, the training and validation data are merged, such that the final model trains on all non-test data. No early stopping is applied; instead, we scale the number of training epochs based on the results of the hyperparameter optimization (HPO):

$$T_{\text{final}} = \text{round} \left(T_{\text{hpo}} \times \frac{N_{\text{hpo}}}{N_{\text{final}}} \right), \quad (16)$$

where T_{hpo} is the epoch at which early stopping was conducted for the best HPO trial, N_{hpo} is the training-set size during HPO, and N_{final} is the size of the combined training and validation data. For the final training, we use the AdamW optimizer (Loshchilov and Hutter, 2019) with a linear warm-up schedule (learning rate rises from $\text{lr} \times 10^{-3}$ to lr over 10 epochs) followed by cosine annealing to a minimum of 10^{-6} .

The identity-constrained test partitions are rather small (160 recordings for DogBark, 758 for SoundWel). To yield a reliable performance assessment and reduce the effect of randomness in training, we repeat Stage 2 five times and provide an averaged performance over five seeds. The performance variation is, for instance, caused by random weight initialization, per-step reparameterization noise ($z = \mu_z + \sigma_z \odot \varepsilon$, $\varepsilon \sim \mathcal{N}(0, I)$), and stochastic data augmentation.

4.4. Test-Time Inference and Segment Aggregation

At test time, audio recordings are processed at their full length rather than as single random crops, ensuring that all temporal information is utilized for classification. As shown in Fig. 2, each recording is partitioned into overlapping segments of 400 ms duration with 50% overlap, producing several chunks. Recordings shorter than 400 ms are represented by a single chunk and right-padded with zeros to the target length.

Each chunk is independently processed by the model to produce a per-chunk prediction, and the final recording-level prediction is determined by a majority vote across all chunks. We used this strategy since the overlapping chunks ensure comprehensive coverage of the recording, and the majority vote is robust to individual chunks containing noise or ambiguous features.

4.5. Implementation Details

All experiments were implemented with PyTorch Lightning and trained on NVIDIA RTX 6000, with 48GB VRAM using Distributed Data Parallelism and 16-bit mixed-precision arithmetic. All hyperparameters are HPO-selected independently per model variant and dataset; exact values are recorded in the experiment configuration files released with the code. Both the waveform and spectrogram encoders are trained from random initialization; no pretrained weights are used. All variants share identical backbone architectures and training conditions to ensure that performance differences reflect the fusion strategy rather than encoder capacity. The **Concatenation** baseline uses the same dual ResNet-18 encoders and mean-pooling aggregation as the proposed UAF model; the two 512-dimensional utterance embeddings are concatenated into a 1,024-dimensional vector and passed to a linear classification head trained with cross-entropy only. The waveform-only and spectrogram-only baselines likewise use single-branch cross-entropy. The UAF uncertainty heads ($W_{\mu}^{(m)}$ and $W_{\sigma}^{(m)}$ in Eq. 3) add approximately 262K model parameters over the Concatenation baseline when $D_z = 128$. All code is publicly available on GitHub at <https://github.com/smAIL-WS/Uncertainty-Aware-Crossmodal-Fusion-for-Classification-of-Animal-Behavior>

5. Results and Discussion

We first provide a quantitative performance comparison of our proposed UAF framework on both benchmark datasets. Then, we further analyze the per-class results, the prediction calibration and confidence, as well as the fusion

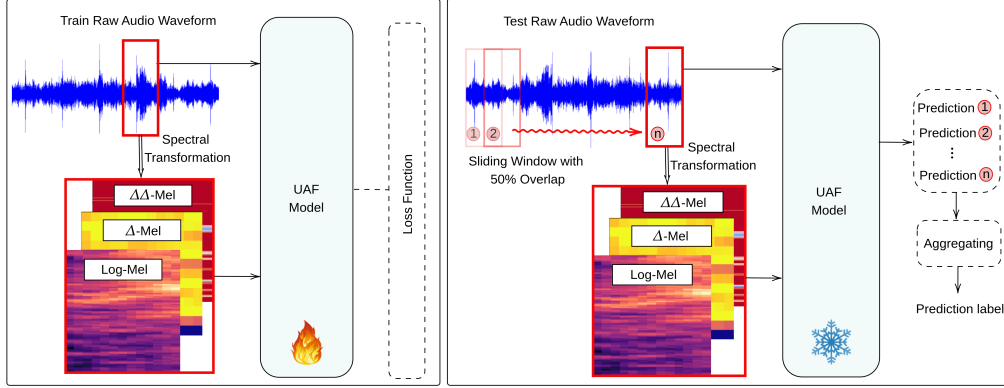


Figure 2: Training and testing procedures. Left: during training, random 400 ms crops are extracted from each recording. Right: during testing, overlapping 400 ms sliding windows with 50% overlap are applied, and predictions are aggregated across all windows.

weight adaptation if one of the representations is corrupted or missing. Model performance is evaluated on held-out test data, with all metrics reported as mean and standard deviation over five independent training runs.

5.1. Quantitative Performance Comparison

In Table 1, we compare seven model variants: two unimodal baselines ((i) waveform-only, (ii) spectrogram-only), (iii) simple concatenation fusion, and four uncertainty-aware fusion (UAF) configurations differing in their temporal module ((iv) mean pooling, (v) GRU, (vi) Transformer, (vii) attention pooling).

On the SoundWel dataset, the spectrogram-only baseline (57.7% acc., 37.2% F1) substantially outperforms the waveform-only baseline (52.2% acc., 34.7% F1), confirming that frequency-domain representations capture more discriminative information for the pig vocalization context. Further, simple concatenation (56.4% acc., 34.3% F1) underperforms the spectrogram-only on both metrics. Hence, merging a weaker waveform stream with a stronger spectrogram without accounting for their relative reliability can dilute the discriminative frequency features. All UAF variants avoid this degradation, which means the uncertainty weighting shifts contribution toward whichever modality produces more confident embeddings. UAF (mean pooling) achieves the best macro-averaged F1-score ($39.7\% \pm 1.3\%$), while UAF (attention pooling) achieves the highest accuracy ($60.3\% \pm 2.5\%$). Our experimental results thus indicate that the primary performance advantage on SoundWel is

Model	SoundWel (pig-identity split, 13/17 classes in test)		DogBark (3 classes, dog-identity split)	
	Acc. (%)	Macro F1 (%)	Acc. (%)	Macro F1 (%)
Waveform-only	52.2 $\pm$ 2.0	34.7 $\pm$ 1.2	53.6 $\pm$ 2.2	44.0 $\pm$ 11.2
Spectrogram-only	57.7 $\pm$ 2.1	37.2 $\pm$ 3.1	69.1 $\pm$ 3.9	63.2 $\pm$ 5.1
Concatenation	56.4 $\pm$ 2.4	34.3 $\pm$ 3.7	62.4 $\pm$ 5.2	59.4 $\pm$ 5.3
UAF (mean pooling)	59.4 $\pm$ 1.9	39.7 $\pm$ 1.3	73.1 $\pm$ 6.9	71.5 $\pm$ 5.7
UAF (GRU)	58.8 $\pm$ 3.7	37.8 $\pm$ 4.2	68.1 $\pm$ 3.5	61.0 $\pm$ 6.2
UAF (Transformer)	57.0 $\pm$ 3.0	38.4 $\pm$ 4.0	63.3 $\pm$ 11.0	51.9 $\pm$ 22.4
UAF (Att. Pooling)	60.3 $\pm$ 2.5	38.0 $\pm$ 1.3	57.0 $\pm$ 6.2	43.0 $\pm$ 10.4

Table 1: Classification performance under identity-based splits. In SoundWel the pig identity inferred from filename, recording year, and recording team). Macro-F1 is computed over the 13 of 17 classes present in the test partition. In DogBark, dogs named Mac and Zoe are considered as the test set. Metrics are mean $\pm$ std over 5 independent training runs. Bold shows the best performing model.

the avoidance of concatenation degradation (-2.9 F1 vs. spectrogram-only) rather than a large absolute gain over the best unimodal baseline. The GRU and Transformer variants are competitive but show larger run-to-run variance, which we attribute to the larger hypothesis space of these models on a dataset of limited size.

On the DogBark dataset, UAF (mean pooling) is the best-performing model across both metrics (73.1% $\pm$ 6.9% acc., 71.5% $\pm$ 5.7% F1), outperforming spectrogram-only by 4.0 percent in accuracy and 8.3 percent in F1-score, and outperforming concatenation by 10.7 and 12.1 percent, respectively. This large margin over concatenation indicates that the adaptive modality weighting of UAF provides a meaningful robustness benefit for small datasets such as DogBark. UAF (Transformer) exhibits unstable training (63.3% $\pm$ 11.0% acc., 51.9% $\pm$ 22.4% F1), with some runs converging well and others collapsing, a known limitation of Transformers when training data per class is limited. Further, UAF (attention pooling) (57.0% $\pm$ 6.2% acc.) underperforms the concatenation method on DogBark, suggesting that the attention pooling aggregation may overfit for DogBark with only six training identities.

Role of temporal architecture. Leveraging a GRU or Transformer component within our UAF framework does not yield consistent performance gains in comparison with a simple mean pooling (Table 1), and increases run-to-run variance. This indicates that local chunk-level uncertainty estimation is sufficient and that introducing sequence dependencies may cause overfitting rather than generalization. We therefore recommend mean pooling as the

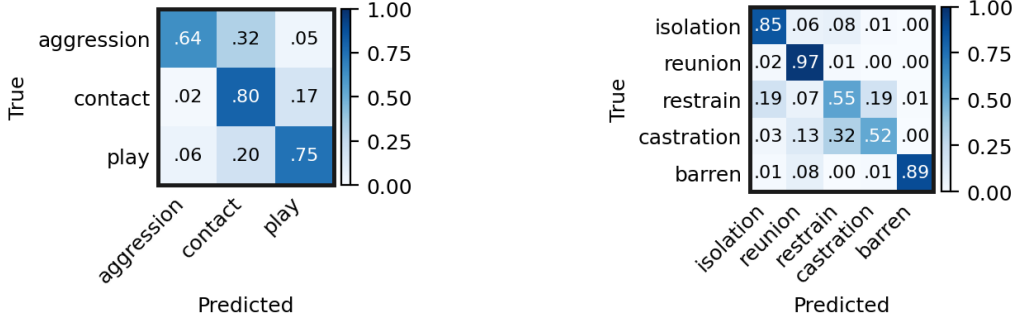


Figure 3: Normalized confusion matrices for UAF (mean pooling) on DogBark (left, 3 classes) and SoundWel (right, 5 welfare-relevant classes). See the full 17-class SoundWel matrix in Appendix A, Fig. A.1.

default temporal module for uncertainty-aware bioacoustic fusion at current dataset scales and focus our results analysis on this variant in the following sections.

5.2. Per-Class Classification Analysis

Figure 3 shows normalized confusion matrices for the UAF (mean pooling) model on both datasets. On DogBark (Fig. 3, left), although there is an imbalance in the number of samples across behavioral classes, recall remains balanced (aggression 0.64, contact 0.80, play 0.75). The largest confusions occur between *aggression* and *contact* barks, both of which are broadband, high-energy temporal calls. For SoundWel (Fig. 3, right), we show five classes that represent both acute and chronic welfare challenges (see Fig. A.1 for the complete 17-class matrix). *Reunion* is classified near-perfectly. *Restrain* in contrast is the most challenging class, frequently confused with *castration* and *isolation*, which is consistent with the overlapping high-stress acoustic profiles of procedural contexts. The contrast between *barren* (well-recognized) and *restrain* (poorly recognized) is noteworthy: *barren*-housing grunts are tonally distinct, whereas *restrain* calls overlap spectrally with other acute-stress vocalizations. This suggests that the primary remaining challenge on SoundWel lies in distinguishing between welfare-negative calls.

5.3. Prediction Calibration

Beyond improving classification accuracy, the uncertainty-aware design aims to produce confidence estimates that are well calibrated. Hence, a

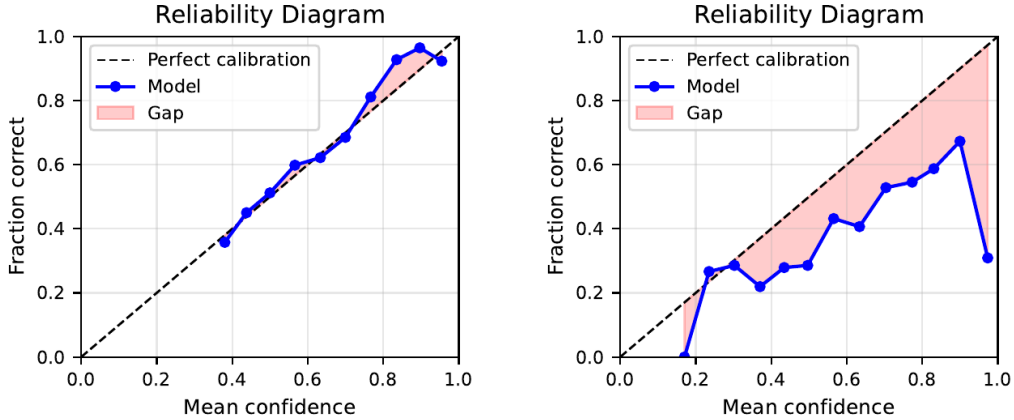


Figure 4: Reliability diagrams for UAF (mean pooling) on DogBark (left) and SoundWel (right). A perfectly calibrated model lies on the dashed diagonal. Points above indicate under-confidence (the model is more accurate than its stated confidence), and points below indicate over-confidence.

prediction reported with, e.g., 90% confidence should be correct roughly 90% of the time. We assess this along several complementary metrics. Figure 4 and 5 visualize this assessment, while Table 2 quantifies it numerically. The Expected Calibration Error (ECE) measures the average gap between stated confidence and actual accuracy, and ECE_{cal} reports this same gap after a simple post-hoc temperature-scaling fix (Guo et al., 2017), showing how much of the miscalibration is easily correctable. The AUROC of the latent uncertainty norm $\|\sigma_z\|$ (Efron, 1979) tests whether the model’s own uncertainty signal can distinguish between correct and incorrect predictions, while the Mann–Whitney U test (Mann and Whitney, 1947) checks whether σ_z is statistically significantly larger for incorrect predictions than for correct ones. The Brier score (Brier, 1950) folds accuracy and calibration into a single proper scoring rule, penalizing confident mistakes most heavily.

The reliability diagrams in Fig. 4 shows that our model is modest about its accuracy for DogBark, while it is overconfident for Soundwel. The reason could be that the DogBark dataset has long calls of the dogs and majority voting over many sliding-window chunks helps to recover the correct label more often than any individual chunk’s own confidence would suggest. On the other hand, Soundwel reflects its greater dataset complexity. However, in Table 2, ECE evaluations show that the reliability for Soundwel could be corrected at test phase by fitting a scalar temperature. Further, the

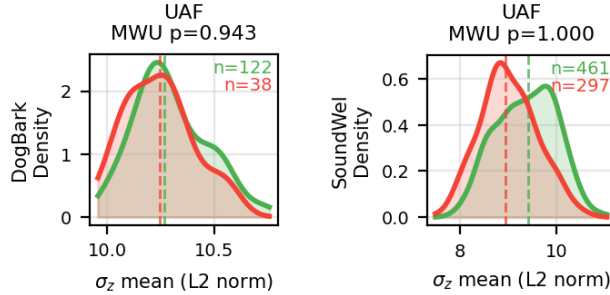


Figure 5: Distribution of $\|\sigma_z\|$ (latent uncertainty norm) for UAF (mean pooling) on DogBark (left) and SoundWel (right). Green: correctly classified recordings; red: incorrectly classified.

Brier scores (0.199 for DogBark, 0.217 for SoundWel) indicate the model’s performance based on accuracy and uncertainty is well below the worst-case value of 1.

The Difficulty-Confidence loss (see Section 3.7) pushes harder samples toward lower values of σ_z . The AUROC (0.415 for DogBark, 0.337 for SoundWel) and p_{MWU} (0.943 for DogBark, 1.000 for SoundWel) values in Table 2 confirm this statement. Consequently, the latent uncertainty could be used as a misclassification signal, such that a low σ_z indicates that the recording is most likely misclassified. For example, see Figure 5, where the distribution of correctly classified samples for the SoundWel dataset has a larger mean. However, this is not a strong misclassification signal for the DogBark dataset, which we attribute to the limited size of this dataset, with only 122 correctly classified and 38 misclassified samples.

Table 2: Uncertainty calibration metrics for UAF (mean pooling).

Dataset	ECE↓	ECE _{cal} ↓	AUROC	Brier _{bin} ↓	p_{MWU}
DogBark	0.195	0.109	0.415 _[0.31,0.52]	0.199	0.943
SoundWel	0.144	0.074	0.337 _[0.30,0.38]	0.217	1.000

5.4. Prediction Confidence by Behavioral Context

Figure 6 shows how confident correct predictions are within each behavioral class: a high median confidence with low spread indicates a class that the model recognizes consistently, while a low median confidence or wide spread

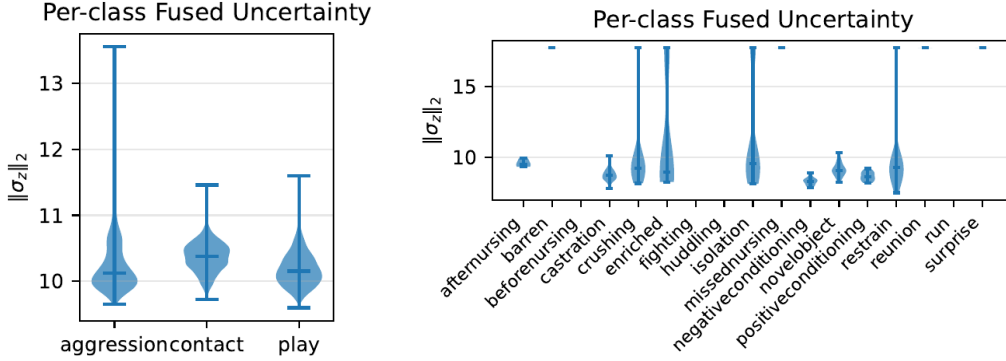


Figure 6: Per-class predicted confidence distributions for UAF (mean pooling) on DogBark (left, 3 classes) and SoundWel (right, 17 classes). Each violin shows the distribution of the model’s maximum softmax probability for correctly classified samples of that class.

indicates a class where the model is unsure between different possibilities, even if it ultimately predicts correctly.

On DogBark, the model produces relatively higher-confidence predictions for aggression barks, which have a distinct broadband impulsive structure, and somewhat lower confidence for contact and play barks. This observation is consistent with the confusion matrix patterns. On SoundWel, confidence distributions are more varied across the 17 classes. These distributions are qualitatively informative for practitioners such that a class with a consistently low-confidence distribution is a candidate for targeted data collection or class-specific fine-tuning.

5.5. Fusion Weight Adaptation Under Targeted Modality Corruption

As an additional robustness experiment, we trained a UAF-based model with Gaussian noise injected into one of the streams during training, including complete signal zeroing to simulate missing inputs and noisy barn environments.

As shown in Fig. 7, in the clean condition (panel a), fusion weights oscillate near 0.5 for both modalities throughout the recording. When Gaussian noise is injected into the waveform at 0.6 s (panel b), the variance weighting shifts to the clean spectrogram. The opposite behavior occurs when the spectrogram is corrupted (panel c) and waveform is the clean input of the model. This indicates that noise injection during training can improve the sensitivity of the fusion weight response in the test phase.

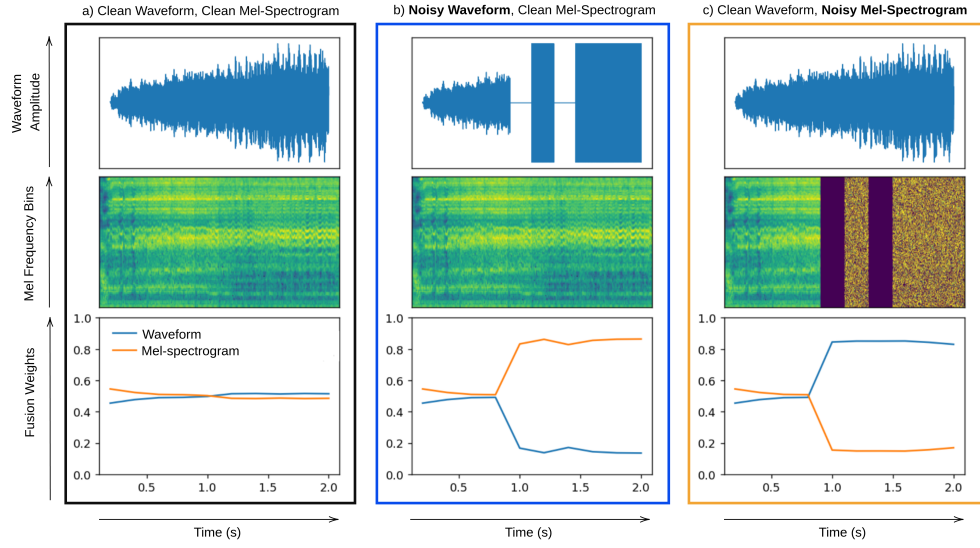


Figure 7: Adaptive modality fusion weights under controlled noise injection for a UAF model trained with deliberative noise augmentation. (a) Clean waveform and mel-spectrogram: fusion weights remain balanced when both signals are clean and equally informative. (b) Clean spectrogram and noisy waveform at 0.6 s: waveform uncertainty rises, shifting fusion weight toward the spectrogram. (c) Clean waveform and noisy spectrogram at 0.6 s: fusion weight shifts toward the waveform.

6. Conclusion

We presented Uncertainty-Aware Fusion (UAF), a dual-stream architecture that estimates per-utterance Gaussian uncertainty for waveform and spectrogram representations and fuses them via dimension-wise uncertainty weighting. Under an evaluation on SoundWel and DogBark that excludes all individuals seen during training, UAF (mean pooling) consistently outperforms both single-modality baselines and static-concatenation fusion. A controlled temporal aggregation ablation shows that simple mean pooling mostly outperforms GRU, Transformer, and attention pooling at current dataset scales. This confirms that the performance improvement arises from uncertainty-driven fusion and not the temporal modeling depth. Calibration analysis of UAF (mean pooling) shows that uncertainty gating reliably improves classification accuracy.

While our approach shows promising results, several directions remain for future research. Both datasets contain a limited number of individuals, resulting in a correspondingly small test pool. Therefore, multi-farm benchmarks with broad environmental variability are needed to characterize generalization at deployment scale. Furthermore, we currently focus on two distinct representations of acoustic monitoring, while the proposed method can be extended to additional sensing modalities. Future methodological work could involve collecting a multimodal dataset and extending UAF to multi-farm settings with complementary modalities, such as video or accelerometry, to address limitations in both scalability and multimodal integration.

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.

CRedit authorship contribution statement

Ehsan Yaghoubi: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Visualization. **Florian Haselbeck:** Conceptualization, Supervision, Writing – review & editing, Funding acquisition.

Data availability

Both datasets used in this study are publicly accessible. The full-size unlabeled DogBark dataset (Yin and McCowan, 2004) is archived at <https://archive.org/details/dog-barks-raw>. We use the annotated version containing 693 samples, which is available at <https://huggingface.co/datasets/cgeorgiaw/animal-sounds>. The SoundWel dataset (Briefer et al., 2022) is provided as a labeled audio database on Zenodo at <https://zenodo.org/records/8252482>. All resources were accessed on 27 July 2026.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

The authors used Google Scholar Labs, Claude Code, ChatGPT, Monica, Grok, Perplexity, and Gemini for improving the clarity of language, brainstorming, literature and dataset search, as well as a critical review of the manuscript. The authors subsequently reviewed and significantly revised all AI-assisted suggestions. The authors assume full responsibility for the content presented in this manuscript.

Acknowledgments

We gratefully acknowledge the support of the Free State of Bavaria for funding the positions of EY and FH via the Bavarian Hightech Agenda. The funder had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

We would like to thank Harshavardan Subramanian and Dr. Georg Back for their constructive discussions and valuable feedback during the development of this work.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M., 2019. Optuna: A next-generation hyperparameter optimization framework, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 2623–2631.

- Baevski, A., Zhou, Y., Mohamed, A., Auli, M., 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations, in: *Advances in Neural Information Processing Systems*, pp. 12449–12460.
- Baltrušaitis, T., Ahuja, C., Morency, L.P., 2019. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 423–443. doi:10.1109/TPAMI.2018.2798607.
- Berckmans, D., 2017. General introduction to precision livestock farming. *Animal Frontiers* 7, 6–11. doi:10.2527/af.2017.0102.
- Bergler, C., Smeele, S.Q., Tyndel, S.A., Barnhill, A., Ortiz, S.T., Kalan, A.K., Cheng, R.X., Brinkløv, S., Osiecka, A.N., Tougaard, J., et al., 2022. Animal-spot enables animal-independent signal detection and classification using deep learning. *Scientific Reports* 12, 21966.
- Bernett, J., Blumenthal, D.B., Grimm, D.G., Haselbeck, F., Joeres, R., Kalinina, O.V., List, M., 2024. Guiding questions to avoid data leakage in biological machine learning applications. *Nature Methods* 21, 1444–1453.
- Bravo Sanchez, F.J., Hossain, M.R., English, N.B., Moore, S.T., 2021. Bioacoustic classification of avian calls from raw sound waveforms with an open-source deep learning architecture. *Scientific Reports* 11, 15733. doi:10.1038/s41598-021-95076-6.
- Briefer, E.F., Sypherd, C.C.R., Linhart, P., Leliveld, L.M., Padilla de La Torre, M., Read, E.R., Guérin, C., Deiss, V., Monestier, C., Rasmussen, J.H., et al., 2022. Classification of pig calls produced from birth to slaughter according to their emotional valence and context of production. *Scientific Reports* 12, 3409.
- Brier, G.W., 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review* 78, 1–3. doi:10.1175/1520-0493(1950)078<0001:V0FEIT>2.0.CO;2.
- Coutant, M., Villain, A.S., Briefer, E.F., 2024. A scoping review of the use of bioacoustics to assess various components of farm animal welfare. *Applied Animal Behaviour Science* 275, 106286.
- Efron, B., 1979. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics* 7, 1–26. doi:10.1214/aos/1176344552.

- Ferrero, M., Chelotti, J.O., Martinez-Rau, L.S., Vignolo, L., Pires, M.F., Galli, J.R., Giovanini, L.L., Rufiner, H.L., 2025. A multi-head deep fusion model for recognition of cattle foraging events using sound and movement signals. *Engineering Applications of Artificial Intelligence* 157. doi:10.1016/j.engappai.2025.013740.
- Gal, Y., Ghahramani, Z., 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning, in: *Proceedings of the 33rd International Conference on Machine Learning*, PMLR. pp. 1050–1059.
- Ghani, B., Denton, T., Kahl, S., Klinck, H., 2023. Global birdsong embeddings enable superior transfer learning for bioacoustic classification. *Scientific Reports* 13, 22876.
- Gómez-Armenta, J.R., Pérez-Espinosa, H., Fernández-Zepeda, J.A., Reyes-Meza, V., 2024. Automatic classification of dog barking using deep learning. *Behavioural Processes* 218, 105028. doi:10.1016/j.beproc.2024.105028.
- Gong, Y., Chung, Y.A., Glass, J., 2021. AST: Audio spectrogram transformer, in: *Proceedings of Interspeech 2021*, ISCA. pp. 571–575. doi:10.21437/Interspeech.2021-698.
- Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q., 2017. On calibration of modern neural networks, in: *Proceedings of the 34th International Conference on Machine Learning*, PMLR. pp. 1321–1330.
- Gupta, G., Kshirsagar, M., Zhong, M., Gholami, S., Ferres, J.L., 2021. Comparing recurrent convolutional neural networks for large scale bird species classification. *Scientific reports* 11, 17085.
- Hagiwara, M., 2023. AVES: Animal vocalization encoder based on self-supervision, in: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE. pp. 1–5. doi:10.1109/ICASSP49357.2023.10095642.
- Han, Z., Zhang, C., Fu, H., Zhou, J.T., 2023. Trusted multi-view classification with dynamic evidential fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 2551–2566. doi:10.1109/TPAMI.2022.3171983.
- He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: *Proceedings of the IEEE Conference on Computer Vision*

- and Pattern Recognition (CVPR), pp. 770–778. doi:10.1109/CVPR.2016.90.
- Hsu, W.N., Bolte, B., Tsai, Y.H.H., Lakhotia, K., Salakhutdinov, R., Mohamed, A., 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29, 3451–3460. doi:10.1109/TASLP.2021.3122291.
- Kendall, A., Gal, Y., 2017. What uncertainties do we need in Bayesian deep learning for computer vision?, in: *Advances in Neural Information Processing Systems*, pp. 5574–5584.
- Kershenbaum, A., Akçay, Ç., Babu-Saheer, L., Barnhill, A., Best, P., Cauzinille, J., Clink, D., Dassow, A., Dufourq, E., Growcott, J., et al., 2025. Automatic detection for bioacoustic research: a practical guide from and for biologists and computer scientists. *Biological Reviews* 100, 620–646.
- Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., Plumbley, M.D., 2020. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28, 2880–2894.
- Li, D., Du, Z., Wang, Q., Wang, J., Du, L., 2024. Recent advances in acoustic technology for aquaculture: A review. *Reviews in Aquaculture* 16, 357–381.
- Liang, Y., Weng, S., Zheng, S., Huang, J.W., 2022. Waveform-logmel dual stream fusion network for sound event detection, in: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE. pp. 1–5. doi:10.1109/ICASSP43922.2022.9978529.
- Loshchilov, I., Hutter, F., 2019. Decoupled weight decay regularization, in: *International Conference on Learning Representations*.
- Mann, H.B., Whitney, D.R., 1947. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics* 18, 50–60. doi:10.1214/aoms/1177730491.
- Manteuffel, G., Puppe, B., Schön, P.C., 2004. Vocalization of farm animals as a measure of welfare. *Applied Animal Behaviour Science* 88, 163–182. doi:10.1016/j.applanim.2004.02.012.

- McLoughlin, M.P., Stewart, R., McElligott, A.G., 2019. Automated bioacoustics: methods in ecology and conservation and their potential for animal welfare monitoring. *Journal of the Royal Society Interface* 16, 20190225.
- Nanni, L., Costa, Y.M., Aguiar, R.L., Mangolin, R.B., Brahnam, S., Silla Jr, C.N., 2020. Ensemble of convolutional neural networks to improve animal audio classification. *EURASIP Journal on Audio, Speech, and Music Processing* 2020, 8.
- Penar, W., Magiera, A., Klocek, C., 2020. Applications of bioacoustics in animal ecology. *Ecological complexity* 43, 100847.
- Peng, X., Wei, Y., Deng, A., Wang, D., Hu, D., 2022. Balanced multi-modal learning via on-the-fly gradient modulation, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8238–8247.
- Schall, E., Kaya, I.I., Debusschere, E., Devos, P., Parcerisas, C., 2024. Deep learning in marine bioacoustics: a benchmark for baleen whale detection. *Remote Sensing in Ecology and Conservation* 10, 642–654.
- Shorten, P.R., Hunter, L., 2023. Acoustic sensors for automated detection of cow vocalization duration and type. *Computers and Electronics in Agriculture* 208, 107760.
- Shorten, P.R., Welten, B.G., 2021. Assessment of a non-invasive acoustic sensor for detecting cattle urination events. *Biosystems Engineering* 207, 177–187.
- Stowell, D., 2022. Computational bioacoustics with deep learning: a review and roadmap. *PeerJ* 10, e13152.
- Stowell, D., Wood, M.D., Pamuła, H., Stylianou, Y., Glotin, H., 2019. Automatic acoustic detection of birds through deep learning: The first bird audio detection challenge. *Methods in Ecology and Evolution* 10, 368–380. doi:10.1111/2041-210X.13103.
- Subedar, M., Krishnan, R., Lopez Meyer, P., Tickoo, O., Huang, J., 2019. Uncertainty-aware audiovisual activity recognition using deep Bayesian variational inference, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6301–6310.

- Takahashi, N., Gygli, M., Pfister, B., Van Gool, L., 2016. Deep convolutional neural networks and data augmentation for acoustic event detection. *arXiv preprint arXiv:1604.07160* .
- Tellamekala, M.K., Amiriparian, S., Schuller, B.W., André, E., Giesbrecht, T., Valstar, M., 2023. Cold fusion: Calibrated and ordinal latent distribution fusion for uncertainty-aware multimodal emotion recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 805–822.
- Tokozume, Y., Harada, T., 2017. Learning environmental sounds with end-to-end convolutional neural network, in: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE. pp. 2721–2725.
- Vu, H., Prabhune, O.C., Raskar, U., Panditharatne, D., Chung, H., Choi, C., Kim, Y., 2024. Mmcows: A multimodal dataset for dairy cattle monitoring. *Advances in Neural Information Processing Systems* 37, 59451–59467.
- Xie, Y., Wang, J., Chen, C., Yin, T., Yang, S., Li, Z., Zhang, Y., Ke, J., Song, L., Gan, L., 2024a. Sound identification of abnormal pig vocalizations: Enhancing livestock welfare monitoring on smart farms. *Information Processing & Management* 61, 103770.
- Xie, Z., Yang, Y., Wang, J., Liu, X., Li, X., 2024b. Trustworthy multimodal fusion for sentiment analysis in ordinal sentiment space. *arXiv preprint arXiv:2404.08923* .
- Yin, S., McCowan, B., 2004. Barking in domestic dogs: context specificity and individual identification. *Animal behaviour* 68, 343–355.
- Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., et al., 2002. The htk book. Cambridge university engineering department 3, 12.

Appendix A. Complete SoundWel Confusion Matrix

Figure A.1 shows the complete 17-class normalized confusion matrix for UAF (mean pooling) on SoundWel, extending the 5-class subset shown in Figure 3.

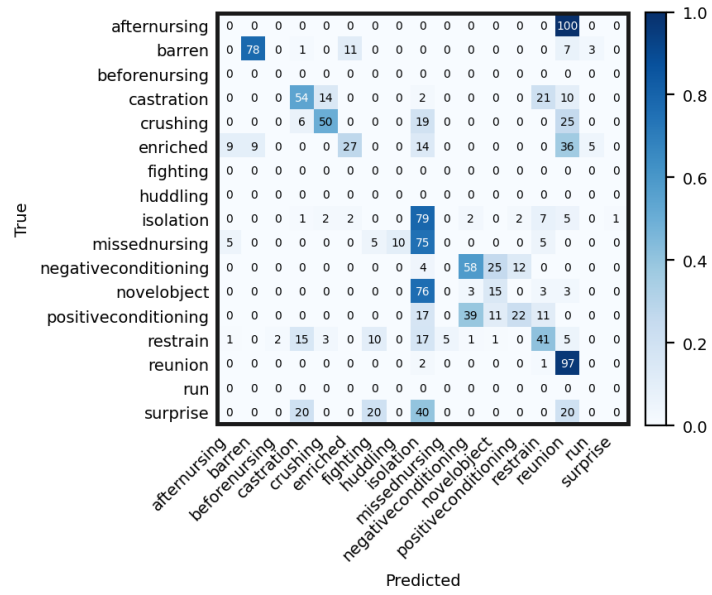


Figure A.1: Complete 17-class normalized confusion matrix for UAF (mean pooling) on SoundWel. Cell values show per-class recall as percent. The four classes absent from the test partition (BeforeNursing, Fighting, Huddling, Run) appear as all-zero rows.

Appendix B. Random Stratified Split Results

Table B.1 reports SoundWel results under a random stratified 80/20 split, in which no identity constraint is imposed and individual animals may appear in both training and test sets. This protocol is commonly used in the bioacoustics literature and we did it to perform a comparison with prior published results such as Briefer et al. (2022). However, because individual-specific acoustic signatures can leak from training into the test partition, the resulting metrics overestimate the generalization ability of the model to truly unseen individuals. Under this split, UAF (GRU) achieves 86.0% accuracy and 76.5% macro F1, surpassing both the simple static concatenation baseline and Briefer et al. (2022).

Model	Acc. (%)	Macro F1 (%)
Briefer et al. (2022) (reported)	81.5	81.2
Concatenation	85.5 ± 0.3	71.6 ± 1.3
UAF (GRU)	86.0 ± 0.3	76.5 ± 1.1
UAF (Transformer)	81.8 ± 0.2	65.5 ± 1.0
UAF (Att. Pooling)	85.7 ± 0.3	74.7 ± 1.5

Table B.1: SoundWel results under a *random stratified 80/20 split* (no identity separation). Metrics are mean $\pm$ std over 5 independent runs. Bold: best among the models.