Uncertainty Quantification of Neural Networks for Marine Mammal Bioacoustic Classification

Andy Yu, Rui Qi Chen, Richik Sen

1. Abstract.

The field of marine bioacoustics, particularly classification, is inherently a very uncertain field due to the complexities of underwater acoustics and limitations of data collection, and prior work primarily focused on improving classification accuracy has not properly studied this high uncertainty underlying the entire field. In this paper, we introduce the first attempts at applying uncertainty quantification to marine bioacoustics, specifically for marine mammals and deep learning approaches. Utilizing the comprehensive Watkins Marine Mammal Database, we apply the InceptionNet convolutional neural network as a new classification technique for marine bioacoustics, on top of which we implement three conventional uncertainty quantification techniques for deep learning: Deep Ensembles, Monte Carlo Dropout, and Bayesian Neural Networks (BNNs). These approaches were then evaluated use several different metrics to assess model performance at both classification accuracy and uncertainty quantification. InceptionNet achieved results comparable to similar convolutional neural networks with an 84% accuracy, while the integration of all three uncertainty quantification methods resulted in notable improvements both to classification accuracy and expected calibration error compared to the baseline model. Deep ensembles resulted in the best model calibration with an expected calibration error 0.0588, while BNNs resulted in a 92% classification accuracy.

2. Introduction and Problem Statement.

Marine bioacoustics, or the study sounds emitted by marine wildlife, are a key component of studying marine wildlife. As sound propagates significantly further underwater than any other sensory medium, many animals leverage increased hearing and specialized vocalizations to perform a variety of tasks from communication of critical information to hunting to exploring the physical marine environment [11]. Within the field of bio-acoustics, classification is one of the most fundamental tasks required, as researchers need to be able to ensure that sounds recorded by hydrophone arrays or other equipment belong to certain species before they can further examine other aspects. As it is, this is

However, classification of sounds is inherently difficult, as the ability for sound to propagate far underwater also means that recordings inevitably pick up significant noise from other sources such as ships, offshore platforms, other wildlife, and underwater geological phenomena, among others. When also considering the fact sound can propagate differently in different seabeds and sound gradients that vary with salinity, pressure, and temperature, there is substantial variation and noise within audio samples retrieved from hydrophones [4].

With recent efforts aimed at attempting to apply machine learning techniques to bioacoustic classification, there has been little interest in attempting to ascertain any limitations or bounds of these approaches beyond dataset accuracy or intuition. However, knowing how accurate and confident classification models are enables marine biologists to better trust automated classifiers, help identify species and conditions where manual labeling may necessary to ensure accurate classification, and assist with targeted data collection for species whose current bioacoustics are limited in data or very uncertain.

We therefore aim to close this gap by applying Monte Carlo Dropout, Bayesian Neural Networks, and Deep Ensemble methods to quantify the epistemic uncertainty present with the problem of marine bio-acoustic classification.

3. Literature Review.

3.1. Learning Approaches to Underwater Acoustic Classification. Aslam et al. provide a comprehensive survey of all underwater acoustic classification research using machine learning techniques, including studies dealing with bio-acoustic classification and marine vessel classification as well as further background on the intricacies of the classification problem and the datasets available [1]. For bio-acoustics specifically, a variety of supervised and unsupervised methods has been tried, with various convolutional neural networks (CNNs) being a common and relatively accurate approach.

At the same time, the paper noted major issues with the current state of research in underwater acoustic classification, most notably dataset availability, highly variable model accuracy caused by the complexity of underwater acoustic propagation, and potential for model overfitting based on highly localized data [1]. These issues all emphasize the need for a quantitative assessment of uncertainty. Models using private datasets cannot have their results easily analyzed or further replicated, and many models training and evaluating on publicly available data do so on fairly small sample sizes and low numbers of species classes that may not be generalizable to the need of the scientific community to classify sounds from a far broader set of marine wildlife and hydrophone arrays in a variety of marine environments. We aim to mitigate all the mentioned limitations by using a commonly studied approach - CNNs - trained on a large, publicly available dataset, while generating accurate uncertainty estimates to quantify the epistemic uncertainty arising from the mentioned intricacies of underwater acoustics.

3.2. Ensemble Methods and Uncertainty Quantification. Bayesian Neural Networks (BNNs), Monte Carlo (MC) Dropout, and ensemble methods are prominent techniques for Uncertainty Quantification (UQ) in neural networks, each offering unique approaches to estimate predictive uncertainty. BNNs incorporate uncertainty by treating model parameters as probability distributions, allowing them to capture both data noise and model uncertainty [8]. MC Dropout provides a practical approximation to Bayesian inference by applying dropout during both training and inference, effectively simulating an ensemble of models to estimate uncertainty [2]. Ensemble methods involve training multiple independent models and aggregating their predictions, which enhances robustness and allows for uncertainty estimation based on the diversity of the ensemble's outputs [5]. These techniques have been applied across various domains to improve the reliability and interpretability of neural network predictions, which could be extended to noisy underwater data.

3.3. Inception-Based Architectures in Sound Classification. Recent work in environmental sound classification has demonstrated the effectiveness of deep convolutional neural networks for handling challenging noise conditions. Inception networks leverage parallel convolutional filters of varying sizes to capture multi-scale features [10]. Such architectures can robustly classify environmental sounds even under adverse conditions [9]. These findings motivate our adoption of an Inception-based architecture for marine mammal sound classification, as we extend these methods within a Bayesian framework to also quantify uncertainty.

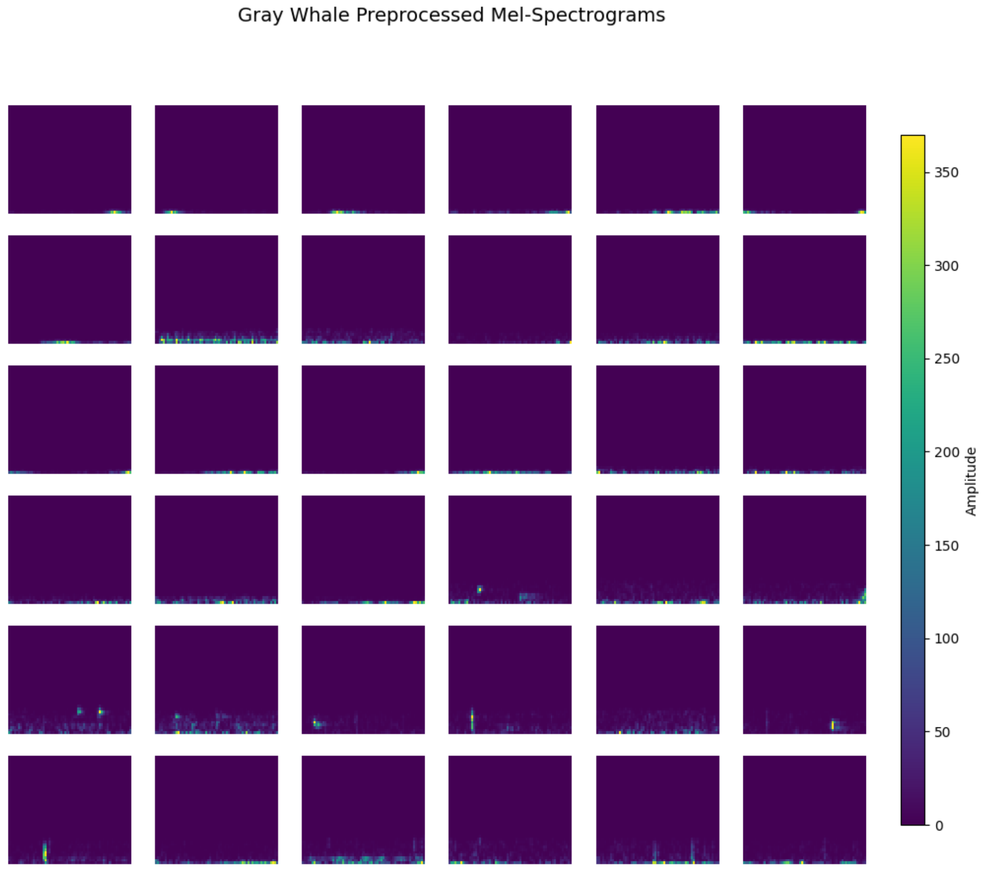

Figure 1. *An diagram of 36 preprocessed mel-spectrograms generated from low-frequency gray whale sound recordings*

4. Dataset and Preprocessing.

We used the Watkins Marine Mammal Database (WMMD) for the training and evaluation of our model architecture. The public dataset comprises of 15,000 manually labeled sound cuts of 51 marine mammals aggregated from multiple databases spanning seven decades [12]. As a result, the Watkins Marine Mammal database is highly variable, with samples from multiple hydrophone arrays in different oceans with a variety sampled frequency bands and track lengths, allowing us to encapsulate the fullest possible extent of the variation and uncertainty associated with classifying marine bioacoustics. These properties should also allow us mitigate the risk of over-fitting on a localized dataset and ensure the reproducibility of our results.

Pre-processing comprised of transforming the sound files into mel-spectrograms, as such an approach is a common approach for CNN bioacoustic classifiers [1]. Specifically, we will apply the preprocessing method adopted by the WhaleNet architecture for the conversion of files into mel-spectrograms [7]. This process comprises of five steps. First is the exclusion of classes with fewer than 100 samples due to concerns over excess data heterogeneity and class imbalance, leaving 26 species out of an original 51 and 14695 retained samples, a retention rate of 94.4% of all samples. Due to differences in hardware, the original dataset possesses sampling rates between 320 Hz and 192 kHz due to a difference in hardware, so all samples are standardized to a uniform sampling rate of 47.6 kHz [12]. Following this step, to standardize the length of each audio recording, the 8000 central timestamps of each audio recording are taken, resulting in a sound cut lasting slightly under 1/6-th of a second. Longer sound recordings are clipped at each end, while shorter sound recordings are padded with zeros on both ends. Next, audio samples are standardized to zero mean and unit variance, then the final mel-spectrogram is generated for these normalized samples. For the generation of the mel-spectrogram, the number of mels, or pitches measured in the mel scale, is set to 41, and the hop-length, or the length of each discretized time segment, is set to 126, yielding a 41 by 64 mel-spectrogram from every recording, where the value at some point (x, y) on the mel-spectrogram is the amplitude of the mel x at time y. As an example, 1 displays a selection of mel-spectrograms from gray whales.

5. Proposed Base Architecture: Inception-Based Network.

5.1. Overview and Rationale.

Marine mammal vocalizations exhibit complex, multi-scale temporal and spectral patterns. This research developed a model capable of processing these features at multiple resolutions simultaneously. The chosen Inception network architecture, as introduced by [10], proved particularly well-suited for this task due to its design that leverages parallel convolutional pathways with filters of various sizes. This multi-scale approach successfully extracted both fine-grained local details and broader contextual features from spectrogram representations of underwater acoustic signals. The implementation of this architecture demonstrated that it was an ideal choice for the marine mammal vocalization classification task.

5.2. Architectural Details.

The implemented architecture adapted the Inception network design for marine mammal vocalization classification. The architecture consisted of the following components:

- **Input Layer:** The network accepted time-frequency representations (log-mel spectrograms) of audio segments extracted from the Watkins Marine Mammal Sound Database. The input consisted of single-channel spectrograms with dimensions of 64 × 41.
- **Initial Convolutional Layers:** Two initial convolutional layers with batch normalization were implemented to extract preliminary features. The first layer used 32 filters, while the second used 64 filters, both with 3×3 kernel sizes and stride of 1.
- **Inception Modules:** Two Inception modules were implemented, each containing parallel convolutional pathways:

  - 1×1 convolutions for capturing pixel-wise features
  - 1×1 convolutions followed by 3×3 convolutions for capturing local patterns
  - 1×1 convolutions followed by two sequential 3×3 convolutions (effectively creating a 5×5 receptive field) for capturing broader features
  - Max pooling operations followed by 1×1 convolutions for preserving important features while reducing dimensionality
- **Intermediate Pooling Layers:** Three max pooling layers were strategically placed: after each initial convolutional block and following the Inception modules. These layers reduced the spatial dimensions of feature maps and introduced translation invariance.
- **Global Average Pooling:** An adaptive average pooling layer reduced the feature maps to 1×1 spatial dimensions, replacing the traditional flattening approach and helping to prevent overfitting.
- **Fully Connected Layers:** The architecture concluded with two fully connected layers. The first layer reduced dimensions from 352 to 256, followed by a final layer that produced class probabilities for different marine mammal vocalization categories. Dropout with probability 0.5 was applied between these layers for regularization.

This architecture successfully balanced model complexity with computational efficiency, enabling effective classification of marine mammal vocalizations while maintaining reasonable training times.

5.3. Integration with Bayesian Inference.

- **Deep Ensembles:** Five independently initialized instances of the Inception network will be trained, and their predictions will be aggregated.
- **Bayesian Neural Network:** Maintaining the same architecture but replacing the final fully connected layer with a Bayesian linear layer to provide uncertainty estimates in predictions.
- **Monte Carlo Dropout:** Retaining dropout layers during inference to quantify uncertainty.

6. Uncertainty Quantification Methods. On top of the baseline Inception Network, we implemented multiple uncertainty quantification methods to enhance the basic softmax probabilities produced by the baseline model. Monte Carlo dropout, BNNs, and deep ensembles were integrated with the Inception-based architecture to provide comprehensive uncertainty estimates.

6.1. Baseline. We were able to quantify uncertainty using a standalone baseline model by utilizing the softmax activation function to output probabilities and confidences.

6.2. Deep Ensemble. Deep ensembles were created by training multiple models with identical architectures but different weight initializations, with ensemble predictions aggregated to obtain both classification results and uncertainty estimates through prediction variance [6].

6.3. Monte Carlo Dropout. For Monte Carlo dropout, the existing dropout layers remained active during inference, and sampling multiple forward passes with different parameters dropped allows the construction of a predictive distribution [3].

6.4. Bayesian Neural Networks. The Bayesian Neural Network approach applied variational inference to the model by replacing the final fully connected layer with a Bayesian linear layer, allowing the network to learn distributions over weights rather than point estimates [8]. This modification provided inherent uncertainty quantification capabilities. These uncertainty quantification methods were systematically evaluated to determine their effectiveness in quantifying model confidence for marine mammal vocalization classification tasks.

7. Experimental Methodology.

The experimental framework began with training a baseline Inception-based model using cross-entropy loss, with Shannon entropy of softmax probabilities serving as the initial uncertainty quantification estimate. The research then integrated three advanced uncertainty quantification methods: Deep Ensembles, MC Dropout, and Bayesian Neural Networks. For Deep Ensembles, five models were trained with different random initializations and bootstrap sampling of the training data to enhance model diversity. The classification accuracy and uncertainty quantification performance of these models were systematically compared. Each method's implementation was evaluated using the comprehensive metrics defined in the evaluation framework, including Expected Calibration Error, mean entropy, and inference time.

7.1. Baseline Model Training & Inference. The baseline model was trained using the AdamW optimizer with a one-cycle learning rate scheduler for efficient convergence. Training was conducted on a V100 GPU with the following hyperparameters:

- **Optimization Algorithm:** AdamW optimizer with initial learning rate of 0.001 and weight decay of 0.01
- **Learning Rate Schedule:** OneCycleLR scheduler with maximum learning rate of 0.01, starting phase of 30%, and cosine annealing for the decreasing phase
- **Loss Function:** Cross-entropy loss with class weights to handle class imbalance
- **Training Duration:** 3000 epochs with batch-wise learning rate updates

7.2. Deep Ensemble Model Training & Inference. The deep ensemble approach involved training five independent models with identical architecture but different random initializations. Each ensemble member followed a training procedure similar to the baseline model, with the following configuration:

- **Ensemble Size:** 5 independent models
- **Training Configuration:** Identical to baseline model (AdamW optimizer, OneCycleLR scheduler)
- **Diversity Enhancement:** Different random weight initializations for each model
- **Training Duration:** 3000 epochs per model

The training process for each ensemble member utilized the same optimization hyperparameters and hardware as the baseline model, including the cross-entropy loss function with class weights, AdamW optimizer with learning rate 0.001 and weight decay 0.01, and OneCycleLR scheduler with maximum learning rate 0.01. In additional, all models where trained using an 80/20 test split, with the random seed standardized to ensure all models has the same train-test split.

Inference was performed by averaging each forward passes performed by each model in the deep ensemble to obtain a predictive distribution.

7.3. Monte Carlo Dropout Training & Inference.

The training configuration for Monte Carlo Dropout was identical to the training configuration for the baseline model. In inference, using an L4 GPU, 100 forward passes were obtained for each sample and then averaged to construct an ensembled predictive distribution.

7.4. Bayesian Neural Network & Inference.

The Bayesian Neural Network approach incorporated uncertainty by modifying the final fully connected layer into a Bayesian linear layer using variational inference. Instead of learning fixed point estimates for weights, the model learned distributions over weights, specifically Gaussian distributions characterized by learnable means and standard deviations.

During training, each forward pass sampled weights from these distributions, allowing the model to capture uncertainty in its parameters. The loss function combined two components: cross-entropy loss, which encouraged correct predictions, and a Kullback-Leibler (KL) divergence term, which regularized the learned weight distributions toward a simple prior. This combination is known as the Evidence Lower Bound (ELBO).

Importantly, dropout regularization was still employed before the final Bayesian layer, with dropout probabilities reduced to 0.2 from 0.5 to mitigate overfitting during training. However, this dropout served purely as a regularization mechanism and was inactive during inference, contributing no uncertainty to the final predictions.

For uncertainty estimation during inference, 30 forward passes (Monte Carlo sampling) were performed through the Bayesian layer, generating a distribution of predictions for each input. The mean and variance of these outputs were used to assess predictive uncertainty.

8. Results & Analysis.

8.1. Evaluation Metrics.

The performance of each uncertainty quantification method was evaluated using comprehensive metrics for both classification accuracy and uncertainty calibration quality. The evaluation framework assessed models using the following metrics:

- **Classification Metrics:** Accuracy, precision, recall, and F1-score to measure predictive performance
- **Uncertainty Calibration:** Expected Calibration Error (ECE) to evaluate the reliability of confidence estimates

The ECE metric specifically measured how well the predicted probabilities aligned with actual accuracy, with lower values indicating better calibration. The evaluation process involved computing these metrics on both training and test datasets, with particular emphasis on test set performance to assess generalization capabilities.

Table 1 presents the comparative performance of all four methods on the test dataset. The results demonstrate that the Bayesian Neural Network achieved the highest classification accuracy (92.51%) among all methods, with the ensemble approach following at 87.82%. The deep ensemble also exhibited the best calibration performance with the lowest ECE (0.0588), indicating more reliable confidence estimates. Monte Carlo dropout provided moderate improvements over the baseline, while all uncertainty quantification methods showed enhanced performance compared to the baseline model across most metrics.

These results confirm that all three uncertainty quantification methods successfully improved both classification performance and confidence calibration over the baseline model,

Table 1

Performance comparison of uncertainty quantification methods on test dataset

Method	Accuracy	Precision	Recall	F1 Score	ECE
Baseline	0.8445	0.8439	0.8445	0.8430	0.1177
Ensemble	0.8782	0.8765	0.8782	0.8765	0.0588
MC Dropout	0.8602	0.8621	0.8602	0.8599	0.1005
BNN	0.9251	0.9280	0.9251	0.9256	0.0921

with the BNN demonstrating the best overall accuracy and the ensemble method achieving superior calibration as indicated by the lowest ECE value.

8.2. Expected Calibration Error Calculation.

The ECE quantifies how well a model's predicted probabilities align with its actual accuracy. ECE measures the difference between confidence and accuracy across different confidence bins. The calculation involves:

1. Dividing predictions into 10 bins based on confidence levels (maximum softmax probability)
2. Computing the average confidence and accuracy within each bin
3. Measuring the weighted absolute difference between confidence and accuracy across all bins

Mathematically, ECE is defined as:

$$ECE = \sum_{b=1}^{N} \frac{|B_b|}{n} |acc(B_b) - conf(B_b)| \tag{8.1}$$

where $|B_b|$ is the number of samples in bin b, n is the total number of samples, $acc(B_b)$ is the accuracy in bin b, and $conf(B_b)$ is the average confidence in bin b.

Figures 2, 3, 4, and 5 in Appendix A present the calibration curves for all four methods, comparing predicted confidence against observed accuracy. A perfectly calibrated model would follow the diagonal line, where confidence equals accuracy, while any curve above and below the diagonal is underconfident and overconfident respectively. As these figures demonstrate, the baseline model was substantially overconfident compared to all three models, and as the ECE decreased, models generally tended to be less confident. For instance, MC Dropout is mostly overconfident, albeit to a lesser degree than the baseline, BNNs are only overconfident for higher-accuracy predictions, and Deep Ensembles were mostly under-confident.

8.3. Performance.

The ensemble approach required approximately five times the training time of other methods due to sequential training of five models. In addition, the Bayesian Neural Network demonstrated much faster convergence than the baseline network, resulting in high accuracy and predictive improvements despite training only for a third of the epochs as the baseline models.

However, during inference, the ensemble demonstrated efficient performance, requiring only five forward passes plus the overhead of loading model weights from disk. In contrast, MC dropout and BNN inference required 30 and 100 forward passes per prediction respectively, resulting in longer inference times compared to the ensemble approach.

Due to the difference in hardware used, as well as non-standardized hyperparameters, a quantitative performance comparison is not possible at this time.

9. Discussion.

9.1. Contributions. In this paper, we successfully applied and compared multiple deep learning uncertainty quantification methods to the problem of marine bioacoustic classification. All three of our approaches to uncertainty quantification proved much more capable at quantifying uncertainty than unmodified models, and also resulted in substantial improvements to classification performance across all four evaluated metrics. In line with prior literature, deep ensembles resulted in the lowest expected calibration error, significantly outcompeting even Monte Carlo Dropout and Bayesian Neural Networks in producing a more accurate predictive distribution than other UQ methods [6]. However, as expected, deep ensembles resulted in significant training costs in training 5 models from scratch, although they proved faster at inference compared to the other two approaches.

The substantial improvement in the predictive accuracy of the Bayesian Neural Network proved to be unexpected, as did the faster convergence rate yielding even lower computational costs than the baseline and MC dropout, despite its worse performance at predicting uncertainty compared to deep ensembles. We believe that the better accuracy and possible cost ratio improvements should be further examined to ascertain the advantage Bayesian Neural Networks may have, and if this is perhaps an application-specific advantage or a more general advantage applicable to all BNNs.

In addition, all four approaches achieved relatively high classification performance and low expected calibration error, indicating that the problem was much less uncertain that we previously imagined. We believe that this may stem from the nature of the WMMD and related datasets, which only comprise of labeled sound-animal pairs that humans were capable of identifying [1, 12]. There may exist much more uncertainty in the broader space of all marine bioacoustic recordings, classified or unclassified.

Overall, uncertainty quantification can provide significant benefits to marine bioacoustic classification. Compared to conventional deep learning networks, our three applied methods not only can model uncertainty, but they also results in improvements to actual predictions, in some cases at minimal additional cost to train and predict. Each method offers various advantages and tradeoffs in computational resources, classification performance, and model calibration, resulting in different use cases depending on what resources and specific tasks marine biologists may be interested in.

9.2. Future Work & Extensions. As mentioned prior, time, resource, and cost restrictions forced the usage of different hardware for each approach, preventing an accurate quantitative performance comparison between the three methods, alongside the use non-uniform hyperparameters. To truly quantify performance and computational cost, further comparative experiments using standardized hardware and hyperparameters would be ideal.

In addition, hyperparameters were largely chosen based on a combination of intuition and past literature and conventions. By applying hyperparameter optimization methods such as Bayesian optimization, genetic algorithms, and particle swarm optimization, among others, model performance can possibly be fine-tuned and improved further to yield even better predictive accuracy.

Although a baseline accuracy of 84.45% compares favorable to many other bioacoustic classification approaches, especially when considering the much broader task our model has to classify 26 unique species instead of the 2-11 classes of prior literature, it is not state-of-the-art [1]. Using a better baseline model, such as WhaleNet, which achieved a 97%+ accuracy on the WMMD, could result in even better accuracy and model calibration to better classify sounds and quantify uncertainty with respect to our prtoblem [7].

Our preprocessing approach, while effective, makes use of a fraction of a second of each sound recording, when some sound recordings can exceed 20 minutes in length [12]. By adjusting our preprocessing approach, we could leverage the entire length of recordings to obtain even more data, exclude less classes while minimizing concerns over data heterogeneity, and consider the depth and complexities of these long recordings instead of classifying based on an extremely short sample.

Finally, bioacoustics is not the only application of underwater acoustic classification. We believe our work could be extended to quantifying the uncertainty other sources of underwater noise if given a sufficient dataset, which could have a wide variety of applications.

REFERENCES

[1] M. A. Aslam, L. Zhang, X. Liu, M. Irfan, Y. Xu, N. Li, P. Zhang, Z. Jiangbin, and L. Yaan, *Underwater sound classification using learning based methods: A review*, Expert Systems with Applications, 255 (2024), p. 124498, https://doi.org/10.1016/j.eswa.2024.124498, https://linkinghub.elsevier.com/retrieve/pii/S0957417424013654 (accessed 2025-03-03).

[2] Y. Gal and Z. Ghahramani, *Dropout as a bayesian approximation: Representing model uncertainty in deep learning*, in international conference on machine learning, PMLR, 2016, pp. 1050–1059.

[3] Y. Gal and Z. Ghahramani, *Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning*, in Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016, M.-F. Balcan and K. Q. Weinberger, eds., vol. 48 of JMLR Workshop and Conference Proceedings, JMLR.org, 2016, pp. 1050–1059, http://proceedings.mlr.press/v48/gal16.html.

[4] W. Kuperman, *Acoustics, Deep Ocean*, in Encyclopedia of Ocean Sciences, Elsevier, 2001, pp. 101–111, https://doi.org/10.1016/B978-012374473-9.00312-X, https://linkinghub.elsevier.com/retrieve/pii/B978012374473900312X (accessed 2025-03-02).

[5] B. Lakshminarayanan, A. Pritzel, and C. Blundell, *Simple and scalable predictive uncertainty estimation using deep ensembles*, Advances in neural information processing systems, 30 (2017).

[6] B. Lakshminarayanan, A. Pritzel, and C. Blundell, *Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles*, in Advances in Neural Information Processing Systems, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds., vol. 30, Curran Associates, Inc., 2017, https://proceedings.neurips.cc/paper_files/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf.

[7] A. Licciardi and D. Carbone, *WhaleNet: A Novel Deep Learning Architecture for Marine Mammals Vocalizations on Watkins Marine Mammal Sound Database*, IEEE Access, 12 (2024), pp. 154182–154194, https://doi.org/10.1109/ACCESS.2024.3482117, https://ieeexplore.ieee.org/document/10720021/ (accessed 2025-04-21).

[8] R. M. Neal, *Bayesian learning for neural networks*, vol. 118, Springer Science & Business Media, 2012.

[9] J. Salamon and J. P. Bello, *Deep convolutional neural networks and data augmentation for environmental sound classification*, IEEE Signal processing letters, 24 (2017), pp. 279–283.

[10] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, *Going deeper with convolutions*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.

[11] P. Tyack, *Bioacoustics*, in Encyclopedia of Ocean Sciences, Elsevier, 2001, pp. 357–363, https://doi.org/10.1016/B978-012374473-9.00436-7, https://linkinghub.elsevier.com/retrieve/pii/B9780123744739004367 (accessed 2025-03-02).

[12] W. Watkins, K. Fristup, M. A. Daher, and Howald, *SOUND Database of Marine Animal Vocalizations Structure and Operations*, Technical Report WHOI-92-31, Woods Hole Oceanographic Institution, Woods Hole, MA, Aug. 1992, https://cis.whoi.edu/science/B/whalesounds/WHOI-92-31.pdf.

Appendix A.
Calibration Curves for the Baseline Model & Uncertainty Quantification Methods.

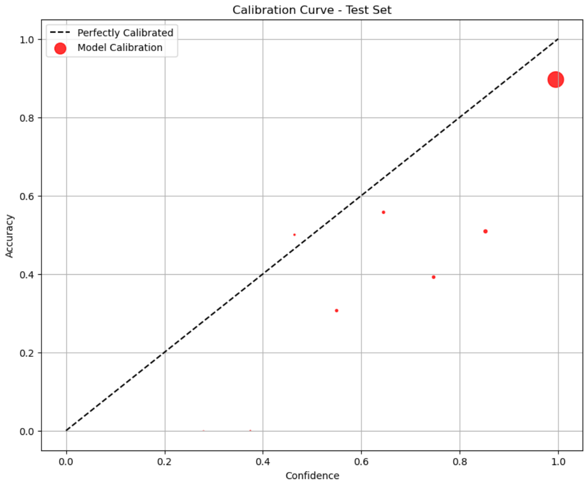

Figure 2. *Calibration curve for baseline InceptionNet*

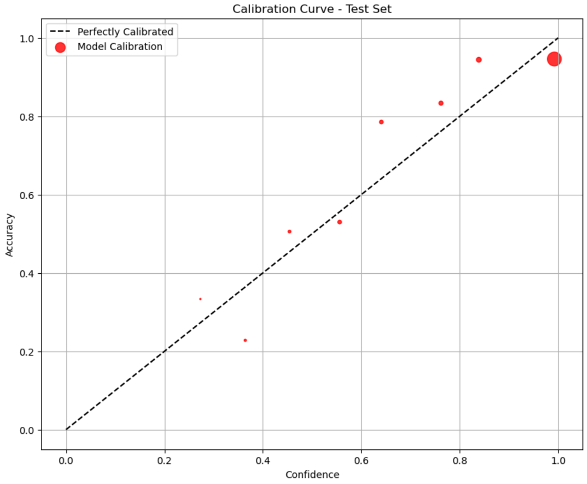

Figure 3. *Calibration curve for Deep Ensembles*

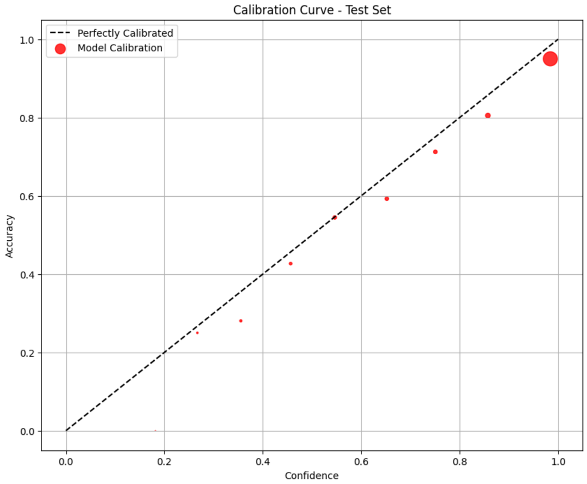

Figure 4. *Calibration curve for MC Dropout*

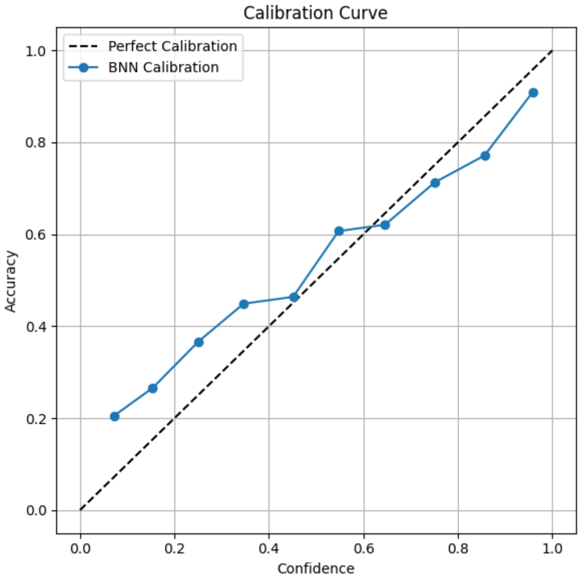

Figure 5. *Calibration curve for BNN*

Appendix B.

Model Implementations. We utilized Georgia Tech's PACE-ICE cluster (V100 GPUs) and Google Colab environments (A100, L4 GPUs) to train and evaluate our models. Training generally took 1 hour for each individual model on the aforementioned GPUs, with deep ensemble training taking approximately 5 hours on PACE-ICE.

Our code is publicly available on GitHub, with the code and further details on installation and execution available at this link: https://github.com/AndyYu25/IUQ_Bioacoustics