Report on Ocean Sound Classification Using Multiple Deep Learning Approaches

1. Introduction

This study explored different deep learning methods for identifying marine species from acoustic recordings.Several types of input representations were tested, including spectrograms, contrastive embeddings, raw audio waveforms, and generated synthetic samples. Each task aimed to understand how various models learn acoustic patterns, how well they separate different species,and how data augmentation affects performance. The goal was to identify which methods provide the most reliable and accurate classification results for ocean sound data.

2. Performance Comparison Across All Tasks

The following table summarises the best accuracies observed for each major approach based on your outputs:

Task / Method	Input Type	Test Accuracy
DNN (Task 1)	Spectrograms	~0.43
CNN (Task 2)	Spectrograms	~0.67 *(best supervised model)*
BiRNN (Task 3)	Spectrograms	~0.54
LSTM (Task 4)	Spectrograms	~0.50
Transformer (Task 5)	Spectrograms	~0.50
Raw Audio 1D-CNN (Task 6)	Raw Waveforms	0.147 *(lowest)*
Contrastive Learning Classifier (Task 7)	Learned Embeddings	~0.56
CNN on VAE-Augmented Data (Task 8)	Spectrograms + VAE	0.6275
CNN on DDPM-Augmented Data (Task 9)	Spectrograms + DDPM	*Not improved; samples lacked structure*

Key observations:

- The **CNN trained on spectrograms** performed the best among standard supervised models.
- The **contrastive learning classifier** reached **~0.56**, outperforming the Transformer, LSTM, and RNN.
- **Raw audio classification** had the weakest performance due to lack of explicit frequency information.
- **VAE augmentation** improved class balance and gave **0.6275**, close to the best performance.
- **DDPM augmentation** did not meaningfully improve accuracy because the generated samples lacked clear acoustic structure.

3. Analysis of Contrastive Learning’s Impact on Class Separation

The tSNE visualization of the contrastive embeddings showed clear improvement in how different species were separated in the feature space. Many classes formed tighter groups compared to earlier supervised models. This suggests the contrastive approach successfully learned patterns that help distinguish between species, even when their raw spectrograms look similar.

Important effects observed:

- **Smaller clusters within each class**, showing the model learned consistent features.
- **Larger distance between different species**,meaning better discrimination.
- Some species that were difficult to classify before now appeared in cleaner, more distinct regions.

The classification report confirms this:

- The contrastive classifier achieved around **0.56 accuracy**,
- Several species reached high precision and recall (some between **0.80–1.00**),
- Overall performance exceeded the RNN, LSTM, Transformer, and raw audio approaches.

Contrastive learning helped the model focus on meaningful relationships in the data. Instead of predicting labels directly,it learned a structured space that separates species more naturally, leading to improved classification results.

4. Quality Assessment of VAE vs Diffusion-Generated Samples

VAE-Generated Samples

The VAE produced spectrograms that:

- captured meaningful class-specific patterns,
- appeared smoother and less detailed,
- were still realistic enough for training,
- helped increase representation for under-sampled species.

After adding VAE samples, the CNN showed:

- **more stable training**,
- better class balance,
- **test accuracy of 0.6275**,
- improved performance especially for classes with fewer examples.

VAE augmentation provided **useful and consistent synthetic data** that enhanced the classification model.

DDPM-Generated Samples

The DDPM model attempted to generate spectrograms through a step-by-step denoising process.However, in this:

- The generated samples looked **highly noisy** and lacked meaningful structure.
- They did not resemble real spectrograms.
- The model had difficulty learning useful details due to having very few samples per class.
- As a result, DDPM augmentation did **not** noticeably improve accuracy.

Although diffusion models can produce high-quality images when trained on large datasets, the current setup did not generate spectrograms suitable for classification.

VAE vs DDPM Summary

Aspect	VAE	DDPM
Spectrogram quality	Clearer, structured, smoother	Mostly noise, lacked shape
Helps minority classes	Yes	Minimal impact
Training stability	Improved	No improvement
Accuracy impact	**Positive (0.6275)**	Not significant
Suitability for this dataset	**Good**	Limited under current conditions

The VAE provided more useful synthetic spectrograms for training, while the DDPM model would likely require more data or a class-conditional design to be effective.

5. Recommendations for Ocean Sound Classification

Based on the results,the following recommendations can guide future improvements:

1. **Spectrogram based CNNs should remain the primary model**,as they consistently achieved the highest accuracy.
2. **Contrastive learning is highly promising** by offering better class separation and strong accuracy without relying on the large labeled datasets.
3. **VAE augmentation should be preferred** when addressing class imbalance or improving minority class performance.
4. **DDPM models may need a larger dataset** or more advanced architectures to produce structured spectrograms suitable for training.
5. **Hybrid architectures combining CNN and Transformer layers** may further improve the performance by capturing both the local and the long-range features.
6. **Feature-space regularization** could provide additional stability and better generalization.
7. **More class-specific data collection** would reduce confusion between acoustically similar species.

6. Conclusion

This study examined multiple deep learning approaches for marine species classification using acoustic data. Models trained on spectrograms performed significantly better than those trained on raw waveforms.Contrastive learning created meaningful feature separations and improved accuracy over several supervised models. Data augmentation using a VAE provided high-quality synthetic samples that improved balance and supported better performance, whereas the diffusion model struggled to produce structured spectrograms and did not enhance classification accuracy.

Combining **spectrogram based CNN models**, **contrastive learning**, and **VAE augmentation** appears to offer the most effective strategy for ocean sound classification under the current dataset conditions.