Experiments in preprocessing techniques for underwater acoustic target recognition

Antriksh Dhand
Supervised by Dr. Dong Yuan

School of Electrical and Computer Engineering
The University of Sydney

February 28th, 2025

Table of Contents

Background

Underwater Acoustic Target Recognition (UATR)

UATR is the analysis of a sonar signal with the aim of determining its source.

Current approaches are manual, with acoustic data being analysed by human sonar operators.

This thesis focuses on the automation of UATR using machine learning techniques, particularly deep learning methods.

Challenges in UATR

The most critical challenge facing UATR research is the scarcity of large, high-quality, labelled datasets.

Table: Volume of common ML audio datasets

Domain	**Dataset**	**Recordings**	**Hours**
Underwater	ShipsEar	90	3
	Deepship	613	47
In-air	FSD50K	51,197	108
	AudioSet	1,789,621	4,971
	Common Voice	13,000,000	18,000

Challenges in UATR

Lack of high-quality, labelled datasets

Why is gathering labelled underwater acoustic data so difficult?

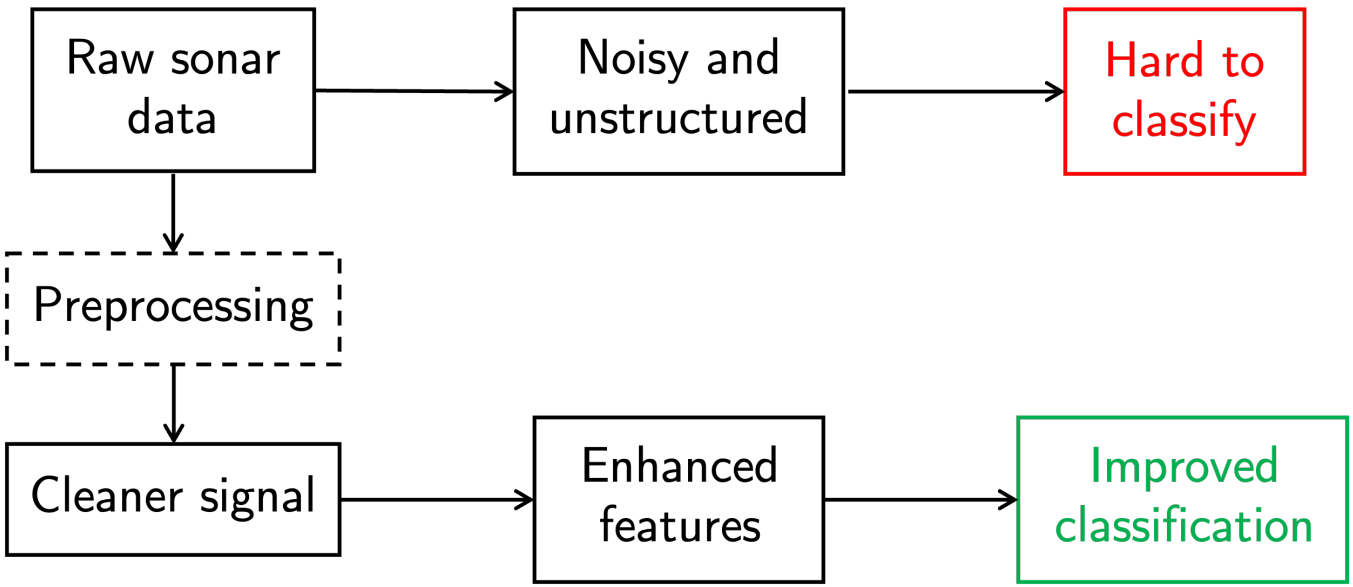
- Resource-intensive and costly: requires time, equipment, maritime logistical coordination, and legal clearance.
- Curation and annotation are highly specialised.
- Many recordings are classified, especially military-related.

Why are underwater recordings prone to being low-quality?

- Signal attenuation (refraction, absorption, scattering, geometrical spreading losses, etc.)
- Multipath propagation leading to delays and signal distortion
- Ambient noise interference interferes with true signal

Motivation

Preprocessing could help UATR work with limited, noisy data.



Aim

This thesis investigates the impact of three preprocessing techniques on UATR classification accuracy using the DeepShip dataset.

Normalisation

Adjusts signal amplitudes for consistency.

Detrending

Removes long-term trends to highlight periodic features.

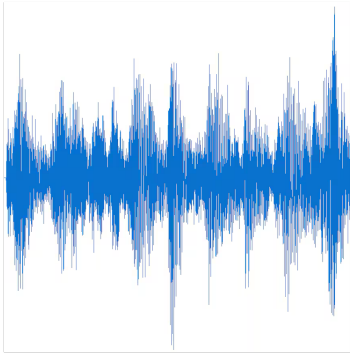
Denoising

Reduces ambient noise for clearer signal interpretation.

Table of Contents

Building a benchmark classifier

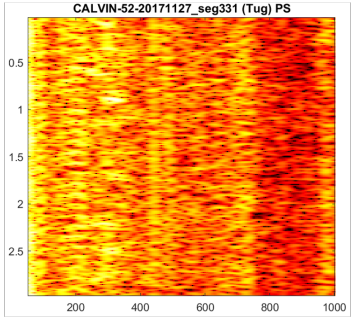
To isolate the effect of each preprocessing technique on signal quality, we first establish a baseline measure of performance with a fixed classification model.



47 hours of DeepShip recordings

Preprocessing

MATLAB



Power spectrogram

Model

TensorFlow

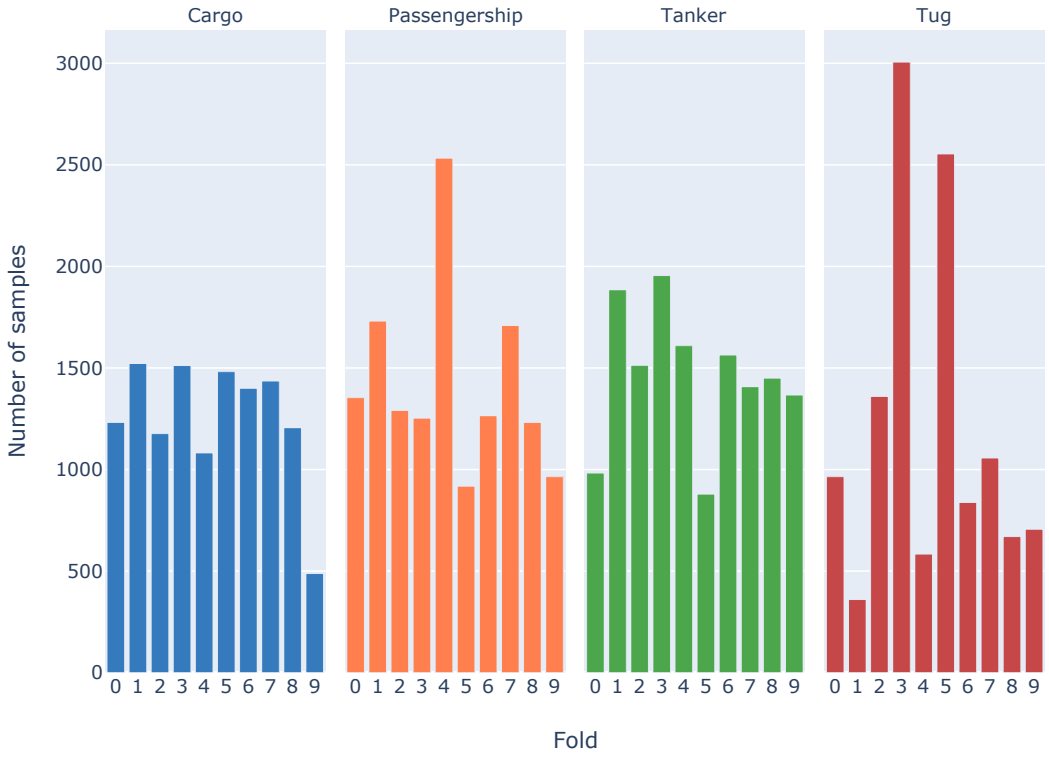
$$\begin{bmatrix} 0.08 \\ 0.22 \\ 0.68 \\ 0.02 \end{bmatrix}$$

Class probabilities

Input data

All experiments used DeepShip, the authoritative benchmark dataset for UATR.

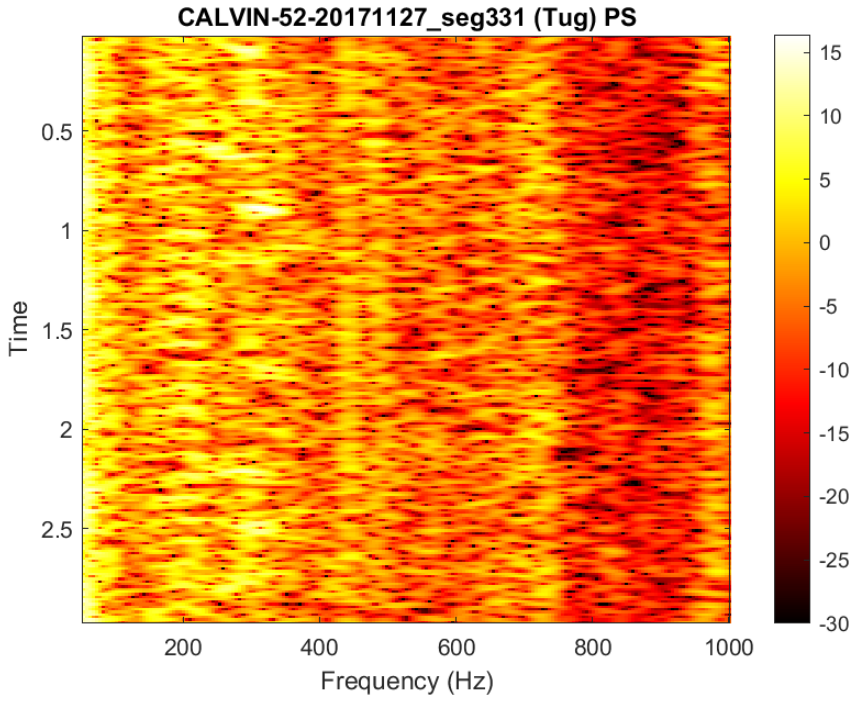
- Recordings were segmented into 53,503 3-second clips to boost training samples and manage computational complexity.
- Segments were divided into 10 folds.



Input features

Power spectrogram

A transformed version of the conventional spectrogram S, calculated as $10 \cdot \log_{10}(|S|^2 + \epsilon)$, designed to enhance interpretability and better align with human auditory perception.



Input features

Processing pipeline

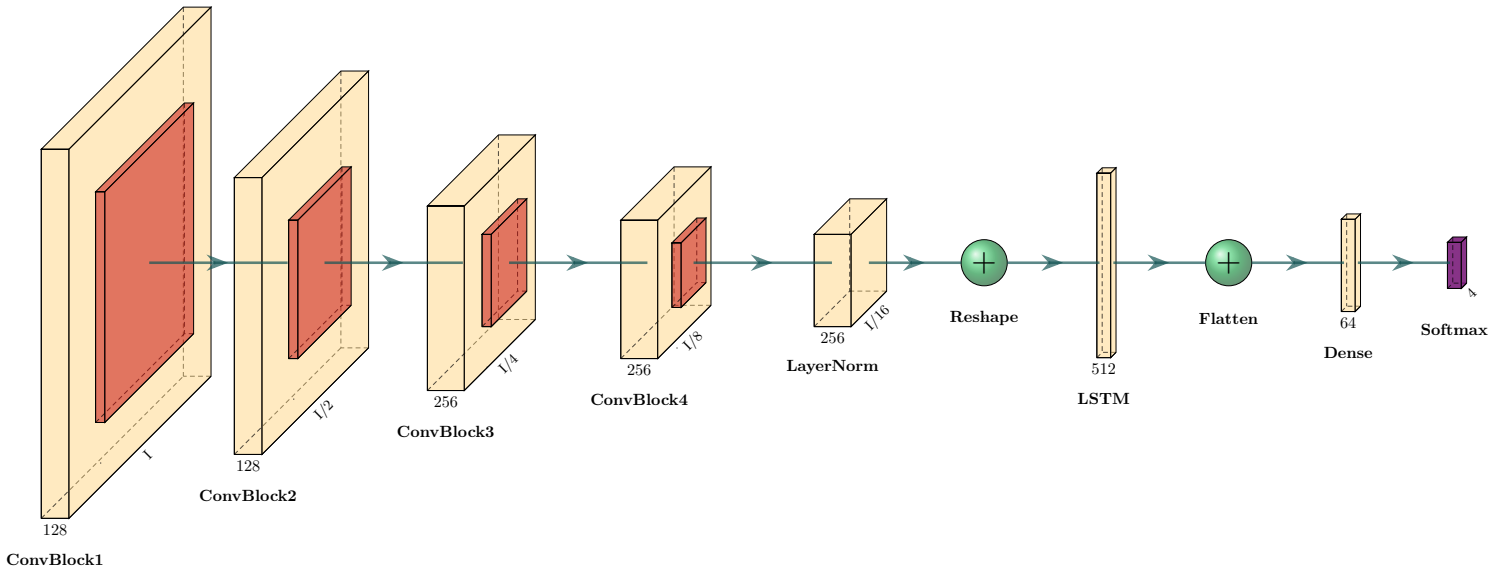
1. Downsample `.wav` file to 5 kHz
2. Perform 0-1 amplitude normalisation
3. Generate spectrogram

Parameter	**Value**
Sampling frequency (`fs`)	5 kHz
Window length (`window`)	40 ms = 200 segments, rounded to 256
Overlap (`noverlap`)	75% of window length
Fast Fourier transform length (`nfft`)	1024 points
Output dimensions	195×231 (frequency bins $\times$ time bins)

4. Amplitude cutoff below -30 dB
5. Frequency cutoff below 50 Hz and 1000 Hz
6. Resize to 192×192
7. Export as `.mat` files

The CNN-LSTM classifier

A hybrid **convolutional neural network–long short-term memory (CNN-LSTM)** classifier was chosen to leverage the strengths of both CNNs for spatial feature extraction and LSTM networks for sequential data processing.



Ultimately, the baseline CNN-LSTM model achieved an overall accuracy of **63.41%** after 5 epochs.

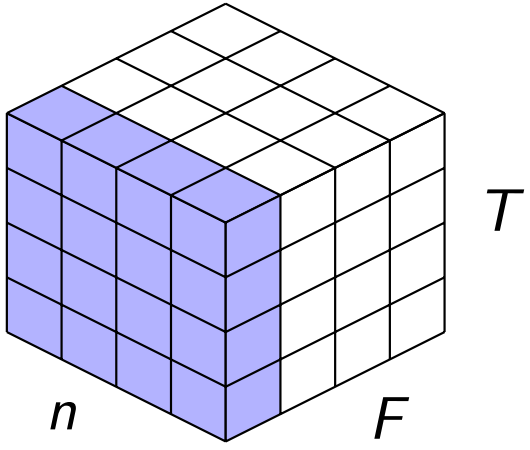
Table of Contents

Table of Contents

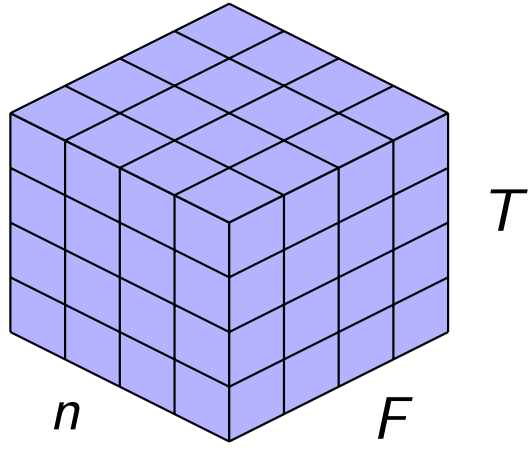
Overview

Hypothesis

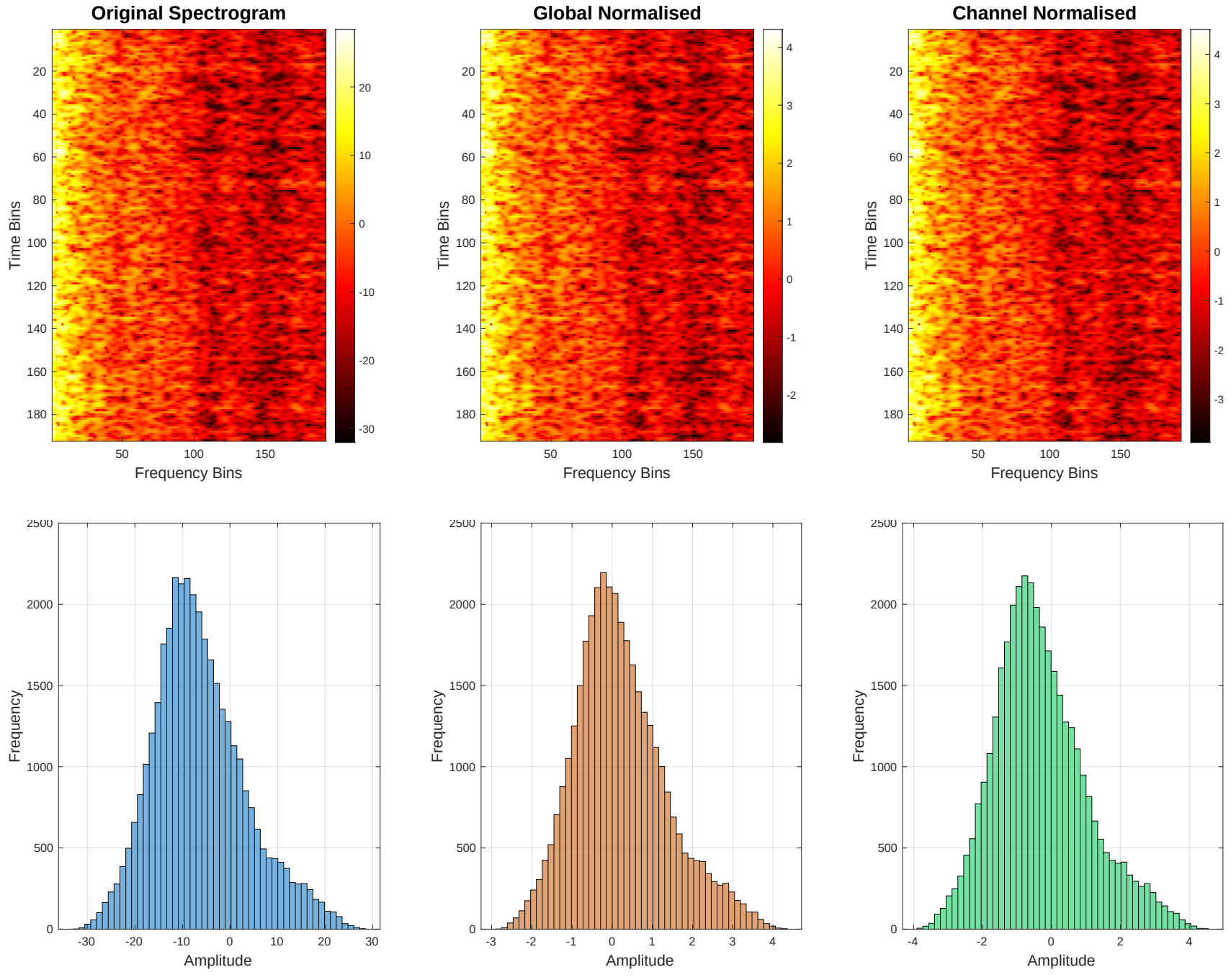
Normalisation improves UATR performance by ensuring all inputs share similar statistical properties, helping maintain a uniform and stable gradient flow throughout the neural network.



(a) Channel-based normalisation



(b) Global normalisation


Original Spectrogram
Global Normalised
Channel Normalised
Time Bins
Frequency Bins
Frequency
Amplitude

Results

There were minimal differences between each normalisation strategy.

Table: Accuracy comparison of normalisation strategies on the DeepShip dataset

Normalisation strategy	**Accuracy (%)**
Global normalisation	62.99
Channel-based normalisation	63.16
Baseline	**63.41**

Discussion

Why didn't normalisation work as expected?

- Power spectrogram may have already standardised the data, reducing the effects of additional normalisation.
- The controlled recording setup (fixed hydrophone) led to uniform feature scales, diminishing the effect of normalisation.

Future work

- Investigate alternative normalisation techniques (e.g. 0-1 normalisation, local normalisation).
- Explore datasets with greater variability (e.g. towed array recordings).

Table of Contents

Overview

Hypothesis

Detrending improves UATR performance by suppressing broadband noise and highlighting transient features important for classification.

The ℓ_1 detrending algorithm modifies the traditional Hodrick-Prescott filter by using an ℓ_1 norm ($\|\cdot\|$) for smoothness:

$$\frac{1}{2}\sum_{t=1}^{n}(y_t - x_t)^2 + \lambda \sum_{t=2}^{n-1} \|x_{t-1} - 2x_t + x_{t+1}\|$$

- λ controls smoothing: higher values detrend more aggressively, lower values retain variability.
- By fine-tuning λ we can strike a balance between removing the trend and preserving the narrowband features critical for classification.

Implementation

- Built on the original MATLAB implementation of the ℓ_1 detrending algorithm.
- Defined regularisation strength λ as a fraction α of λ_{max} (upper bound for detrending).
- Trained CNN-LSTM model on detrended spectrograms using $\alpha = 10^{-2}, 10^{-1}, 1$.

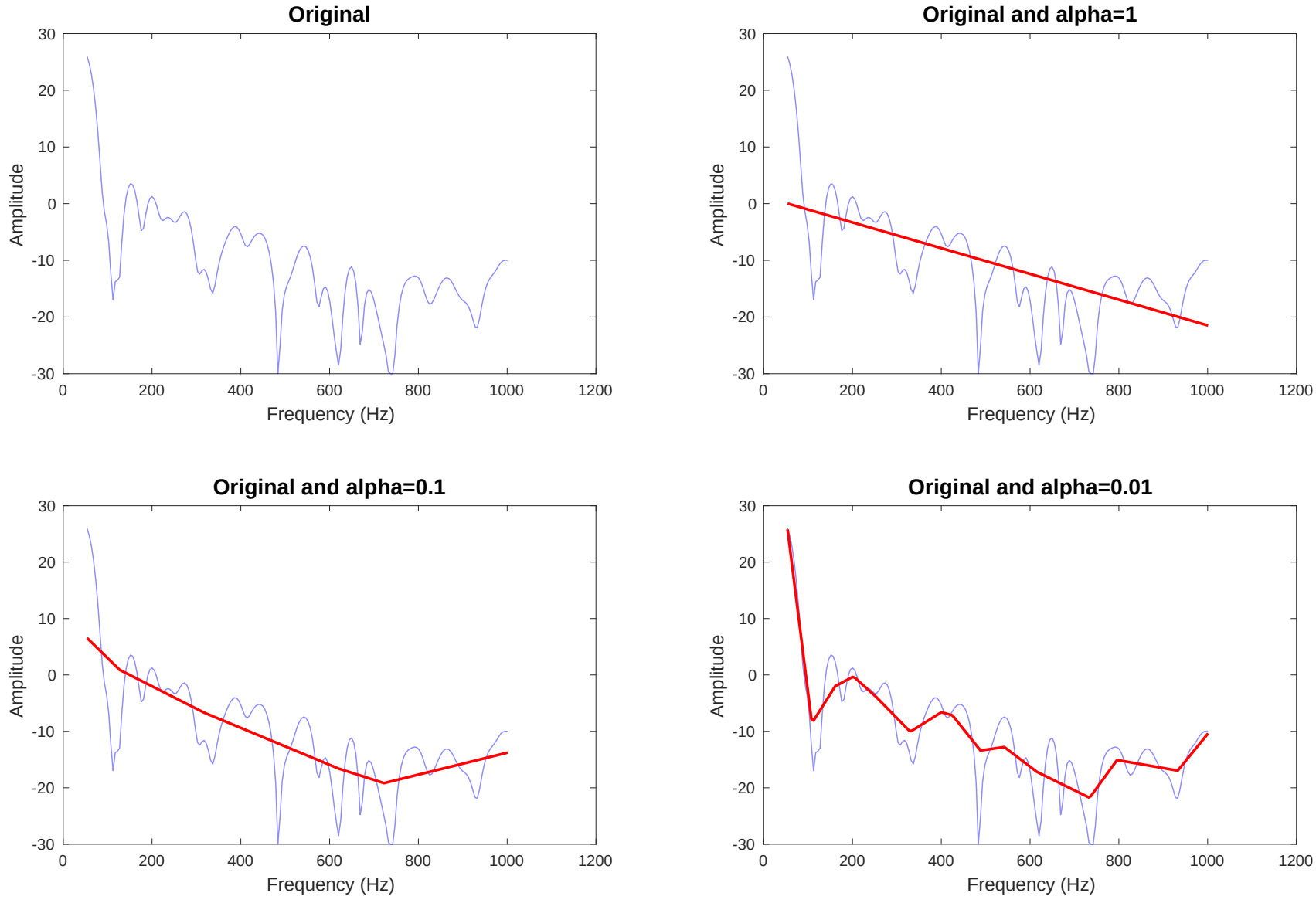


Figure: Overlay of a time segment with its corresponding ℓ_1 trend at various α values

Results

All three detrending configurations resulted in lower classification accuracy compared to the baseline.

Table: Classification results using ℓ_1 detrending algorithm at various α

Detrending parameter	**Accuracy (%)**
$\alpha = 10^{-2}$	48.22
$\alpha = 10^{-1}$	52.03
$\alpha = 1$	55.63
Baseline (no detrending)	**63.41**

Discussion

Why didn't detrending work as expected?

- The short segment duration of 3s may have prevented accurate trend estimation.
- Detrending altered spatial and temporal characteristics, potentially hindering the CNN-LSTM model's learning.
- Low values of α may have been over-smoothing, while high values of α may have been under-suppressing.

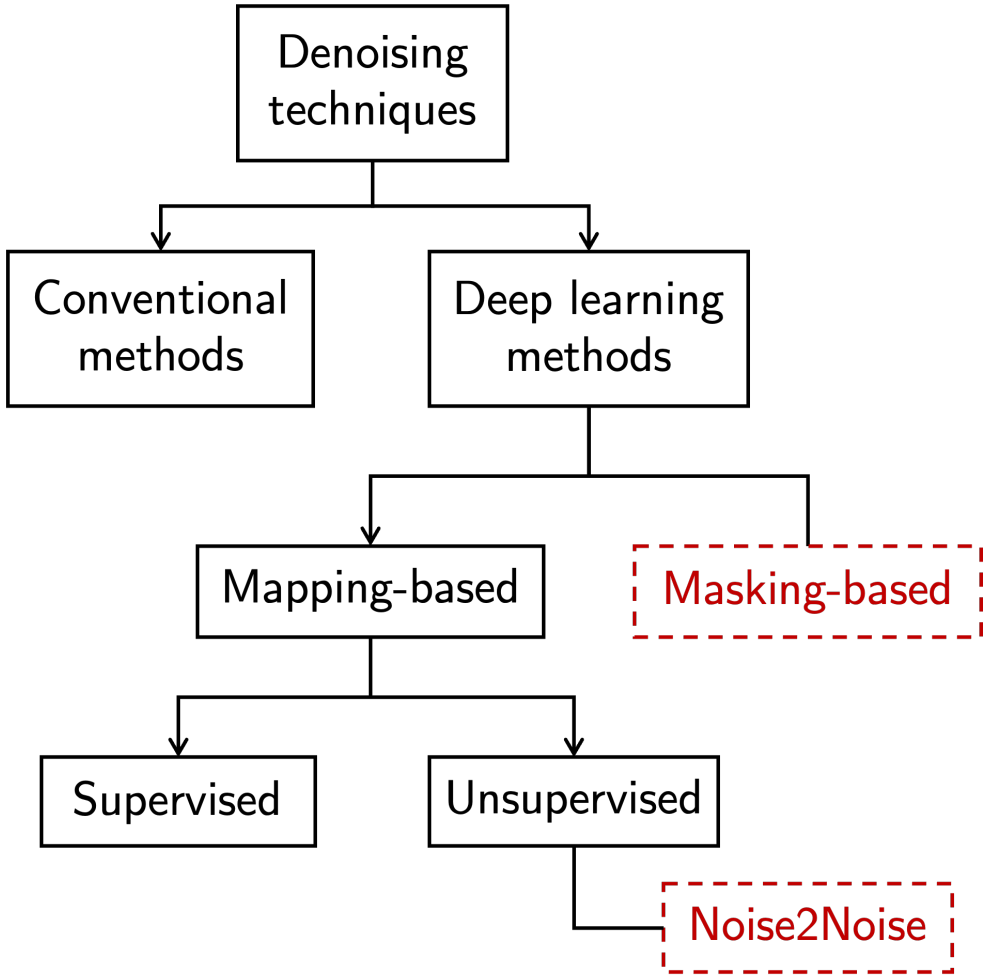
Future work

- Investigate detrending's impact with different model architectures.
- Experiment with longer audio segments to improve trend estimation.
- Compare ℓ_1 detrending to alternative methods (e.g. wavelet-based detrending).

Table of Contents

Overview

This experiment aims to uncover the efficacy of deep learning-based **image denoising techniques** for underwater acoustic spectrograms.



Experiment 1: Unsupervised denoising with Noise2Noise

Noise2Noise (N2N)

The Noise2Noise model can learn to denoise images using only pairs of independently corrupted noisy images provided that

- the noise has a zero-mean distribution, and
- the noise in the input and target images are uncorrelated.

Objectives

1. To validate the N2N approach by recreating its performance on natural images.
2. To investigate its applicability to underwater acoustic spectrograms, a domain where the N2N assumptions are challenging to approximate.

Experiment 1: Unsupervised denoising with Noise2Noise

We compared results between two encoder-decoder architectures: a simple convolutional neural net ("Irfan"), and the well-known U-Net architecture.

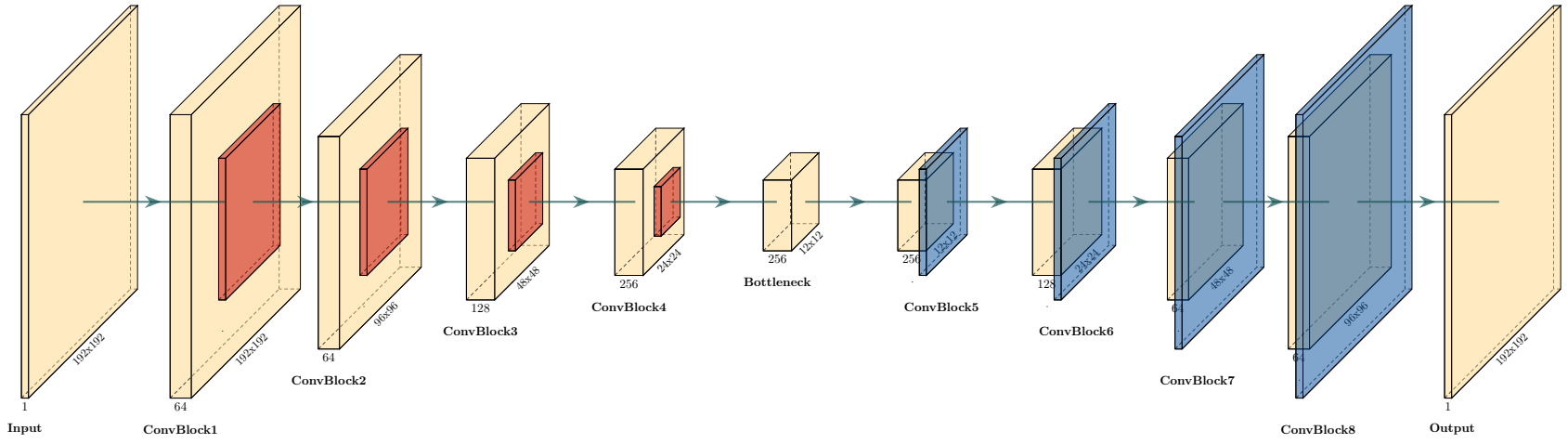


Figure: The Irfan model, adapted from Irfan et al. (2020)

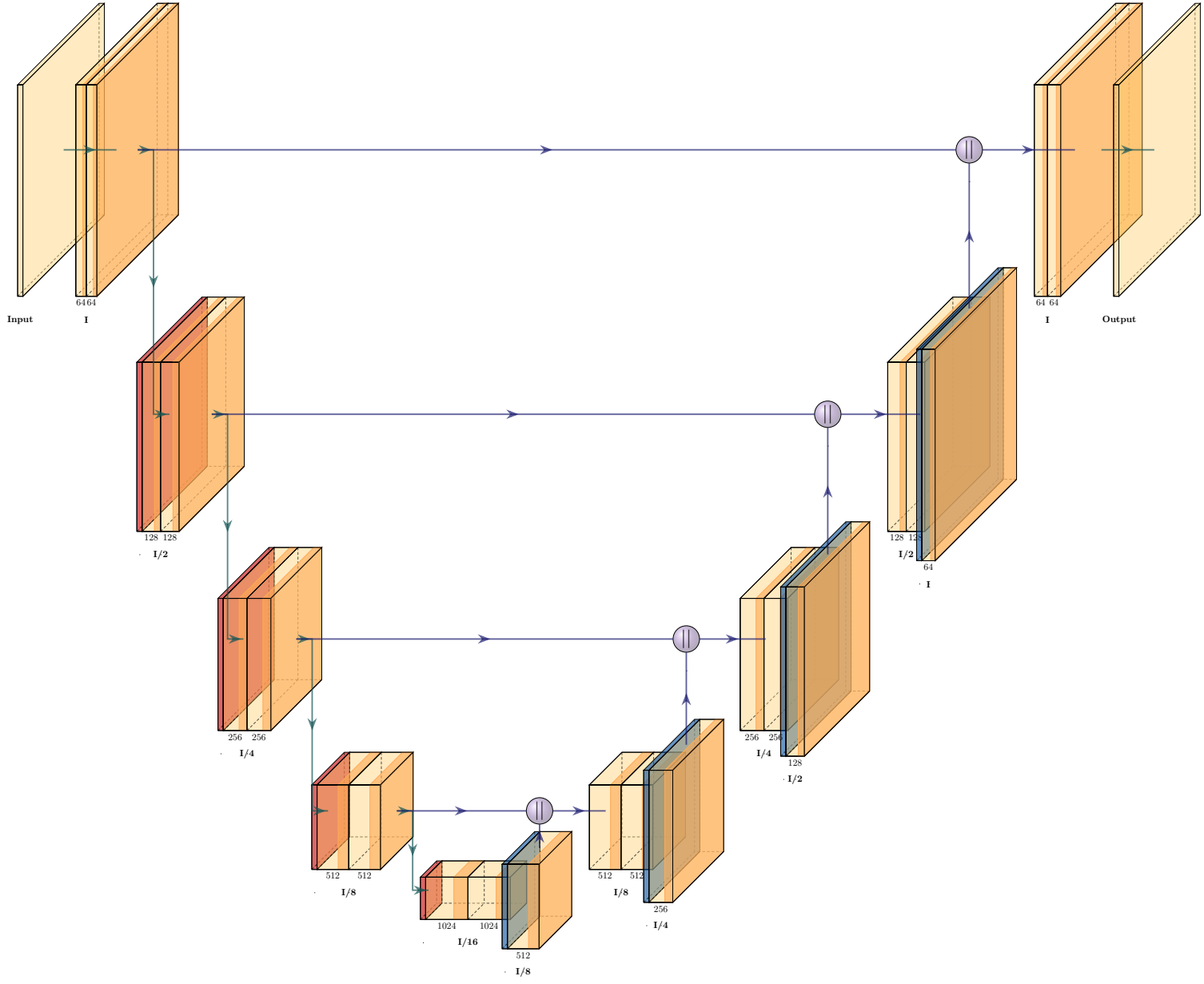
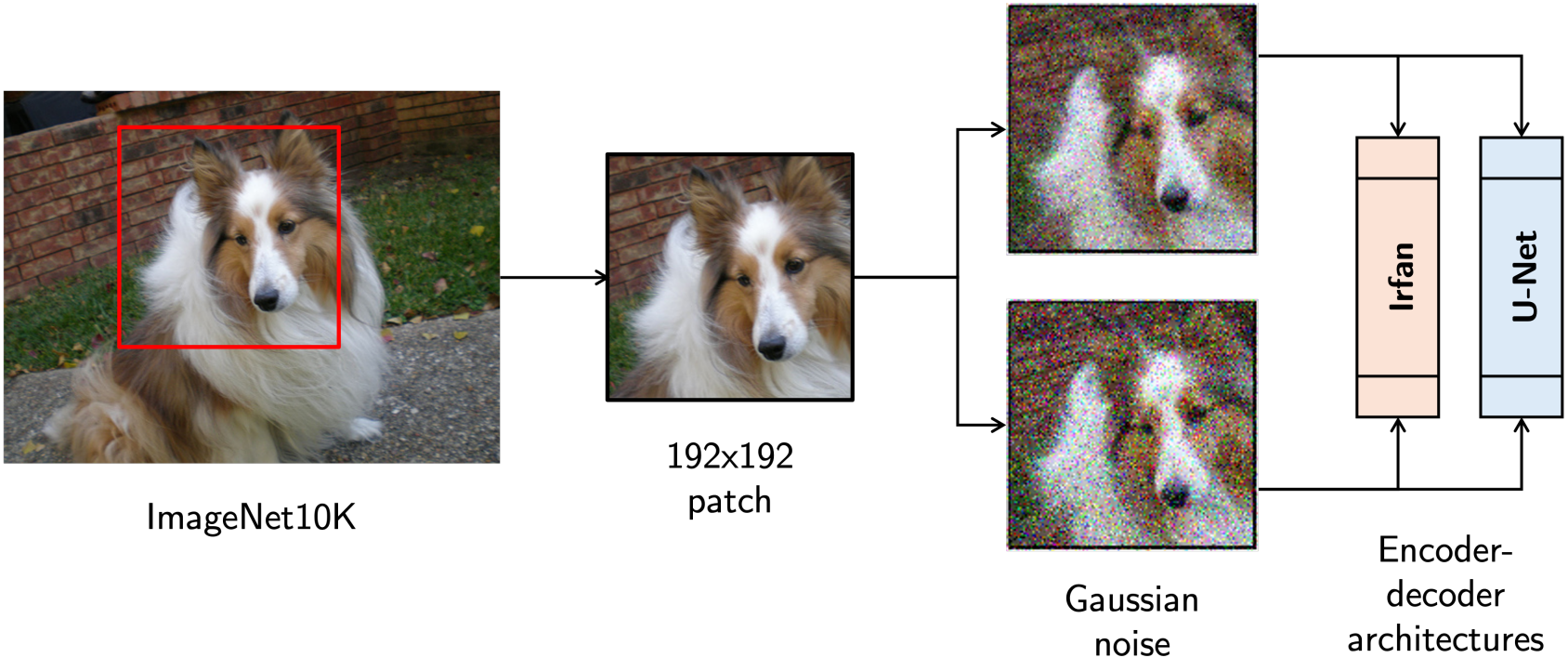


Figure: The U-Net model, first proposed in Ronneberger et al. (2015) for image segmentation

Experiment 1: Unsupervised denoising with Noise2Noise

Part I: Recreating the original paper

In this part we replicate the N2N framework to validate the original methodology and results presented in the seminal 2018 paper.



Experiment 1: Unsupervised denoising with Noise2Noise

Part I: Recreating the original paper

Our U-Net model achieved a peak signal-to-noise ratio (PSNR) of 29.59 dB, only slightly lower than the original paper's 31.06 dB.

Table: Performance of the Irfan and U-Net models on the BSD testing set for our recreation of the Noise2Noise paper.

Strategy	Model	Loss	PSNR (dB)
Supervised	Irfan	0.0094	21.48
	U-Net	0.0017	29.78
Noise2Noise	Irfan	0.0080	22.15
	U-Net	00.18	29.59

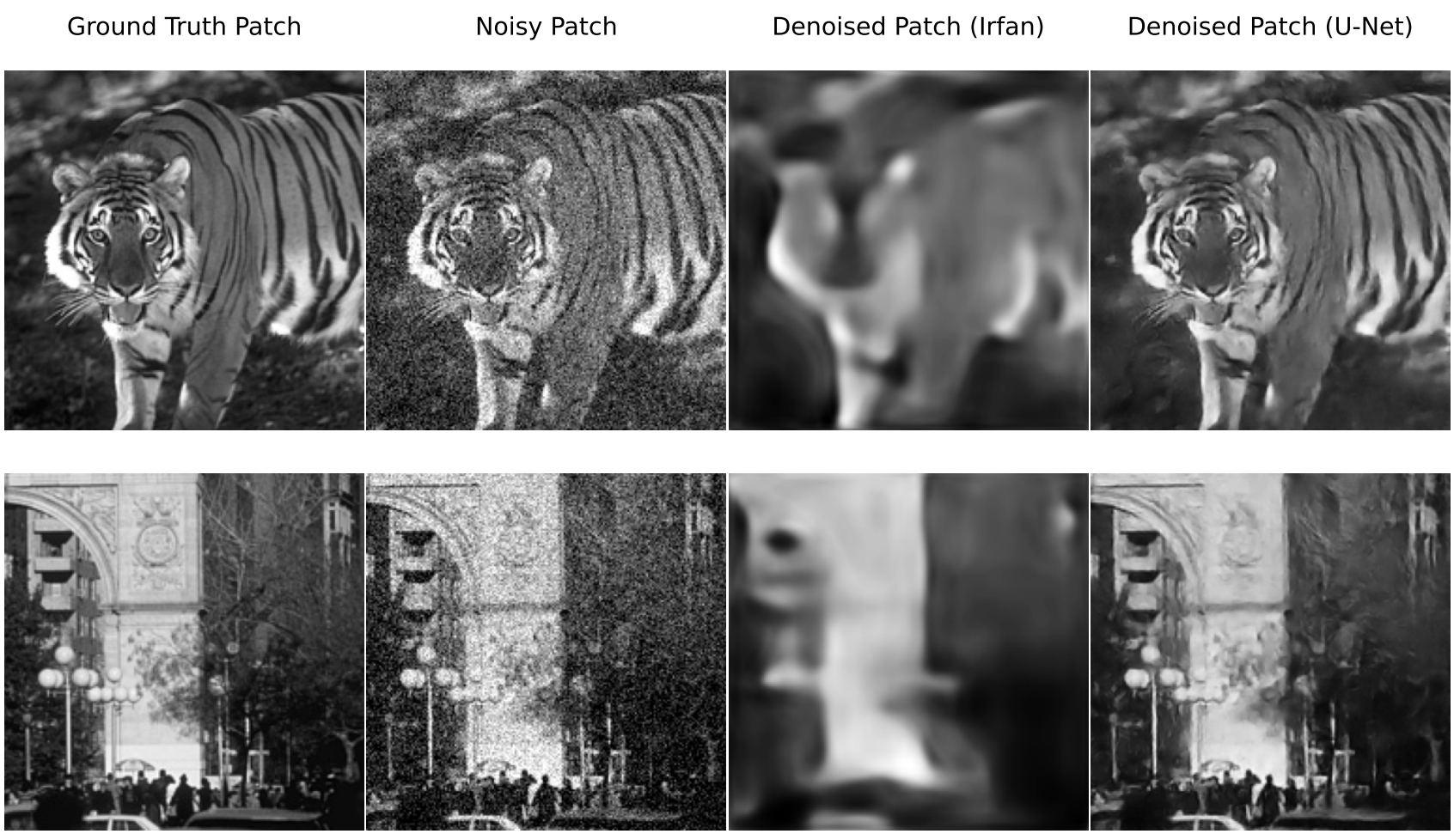


Figure: Comparison of ground truth patches, noisy patches, and denoised outputs

Experiment 1: Unsupervised denoising with Noise2Noise

Part II: Adapting N2N for underwater acoustic spectrograms

Challenge

Noise2Noise requires:

1. Paired recordings of the same event.
2. Independent, zero-mean noise.

But no public hydrophone array datasets exist to satisfy these assumptions.

Can we approximate these assumptions using spectrograms from the same vessel recorded at different times?

Experiment 1: Unsupervised denoising with Noise2Noise

Part II: Adapting N2N for underwater acoustic spectrograms

Implementation

- Filtered DeepShip dataset down to 37,377 spectrograms from vessels with multiple recordings.
- Created custom data generator to randomly pair spectrograms from the same vessel but different recordings.
- Trained Irfan and U-Net for 50 epochs using mean squared error (MSE) and structural similarity index measure (SSIM) as loss metrics.

Experiment 1: Unsupervised denoising with Noise2Noise

Part II: Adapting N2N for underwater acoustic spectrograms

Both models failed to converge during training, resulting in low SSIM scores and blurry, overly-smoothed spectrograms.

Table: Comparison of loss and SSIM values for the Irfan and U-Net models under the Noise2Noise approximation.

Model	Loss	SSIM
Irfan	2.12×10^{-2}	0.118
U-Net	1.77×10^{-2}	0.129

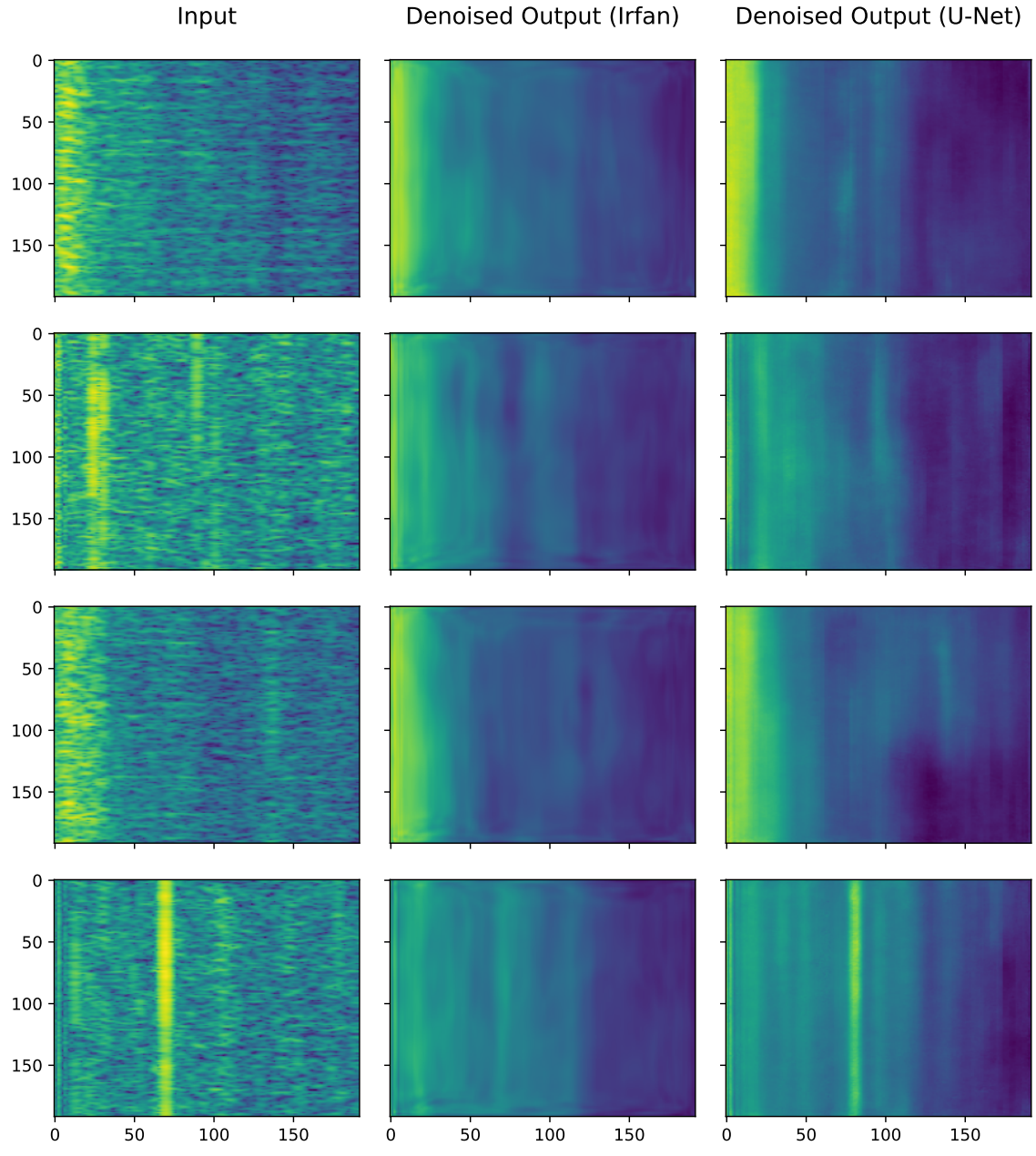


Figure: Outputs generated by the Irfan and U-Net models for the Noise2Noise approximation experiment

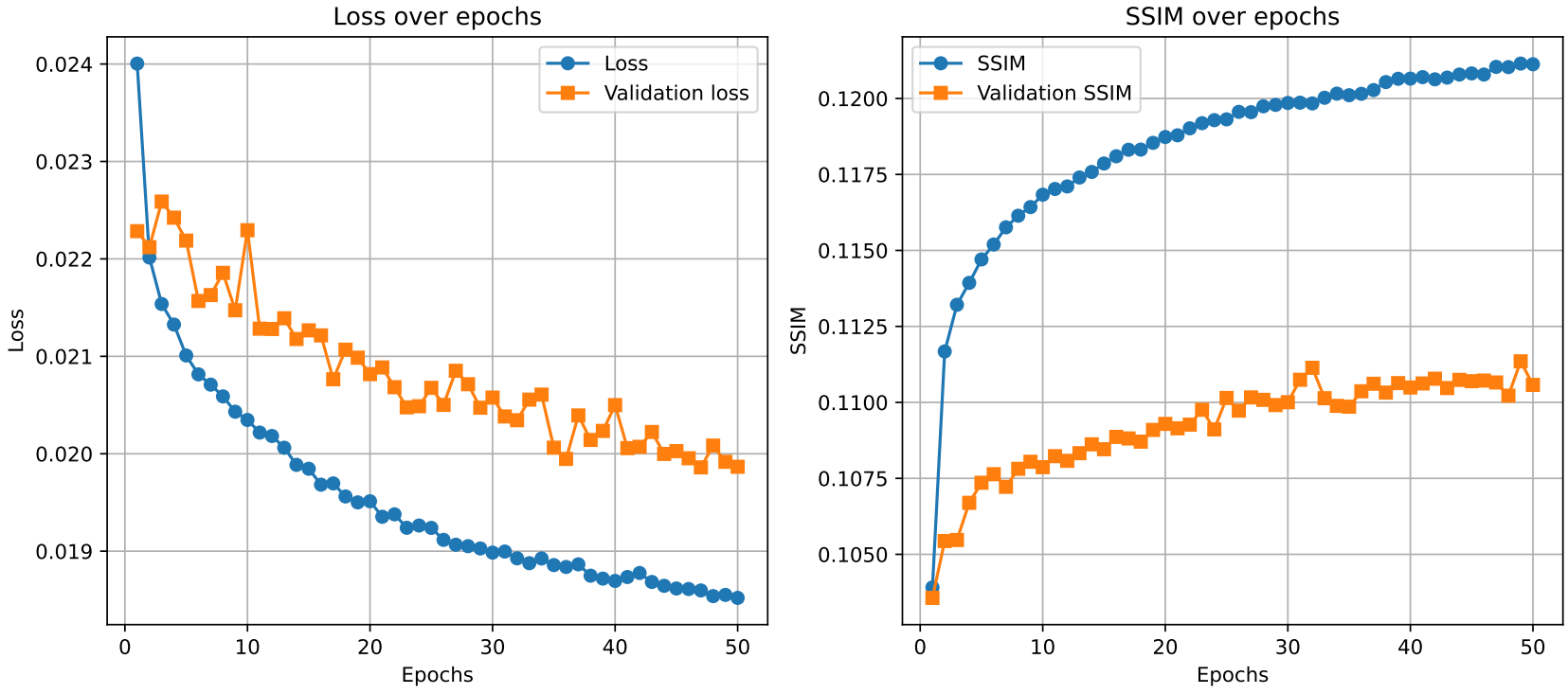


Figure: Loss and SSIM curves for the U-Net model

Experiment 1: Unsupervised denoising with Noise2Noise

Part II: Adapting N2N for underwater acoustic spectrograms

Why didn't N2N work on underwater spectrograms?

The underwater environment is too dynamic and inconsistent (weather, currents, biological activity, etc.) to approximate the N2N assumptions across days or even hours.

Future work

- Collect hydrophone array data for multiple recordings of the same event.
- Explore alternative denoising frameworks better suited to underwater acoustics.

Experiment 2: Masking-based denoising

Overview

Objective

To develop a deep learning model capable of accurately segmenting narrowband events from spectrograms while effectively removing noise.

Implementation

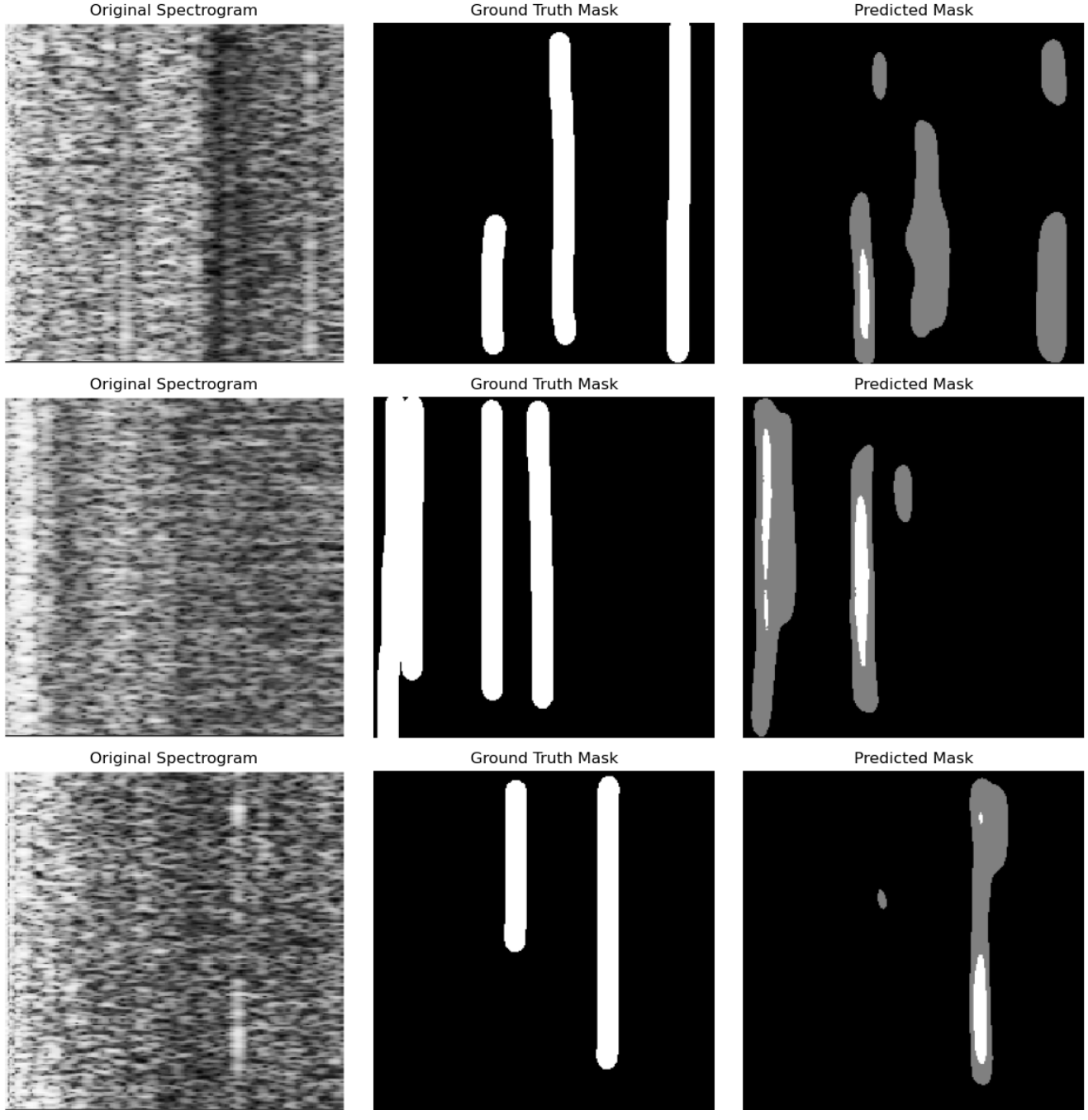
- 356 spectrograms were manually labelled to highlight narrowband frequencies and tracks of interest.
- These were converted into binary masks and divided into an 80-20 train-test split.
- The U-Net model was trained on the data for 500 epochs.

Experiment 2: Masking-based denoising

Results

The model resulted in a high binary accuracy of 93.97% but mediocre binary intersection-over-union of 0.49.

This suggests that the predicted masks were effective at accurately identifying many narrowband events but failed to capture all instances.


Original Spectrogram
Ground Truth Mask
Predicted Mask
Original Spectrogram
Ground Truth Mask
Predicted Mask
Original Spectrogram
Ground Truth Mask
Predicted Mask

Experiment 2: Masking-based denoising

Discussion

Challenges

- Low spectrogram resolution led to limited frequency details.
- Manual annotations conducted by a non-expert may have impacted ground-truth quality.
- The small dataset of 356 samples may have limited model generalisation.

Future work

- Use higher-resolution spectrograms and expert-labelled masks.
- Explore automatic masking techniques for ground-truth annotations, such as HIDE & SEEK.
- Evaluate the impact of generated masks on CNN-LSTM classification performance.

Table of Contents

Summary of results

While normalisation, detrending, and denoising did not improve classification accuracy, these experiments provided key insights for future work:

- **Normalisation** may have a greater impact on datasets with higher variability (e.g. dynamic towed arrays or diverse acoustic environments).
- **Detrending** may be unsuitable for short input segments and classifiers like CNN-LSTM, as it can disrupt key spectral features.
- **Unsupervised mapping-based denoising frameworks** such as N2N struggle in the underwater environment due to difficult-to-approximate assumptions.
- **Masking-based denoising** shows potential, but requires higher-resolution spectrograms and expert-labelled ground-truth masks for better accuracy.

These slides, along with my thesis dissertation and all code written for the experiments discussed today is available at github.com/antrikshdhand/thesis.

Thank You

Any questions?

Contact information

- Email: adha5655@uni.sydney.edu.au
- GitHub: github.com/antrikshdhand
- LinkedIn: linkedin.com/in/antrikshdhand

Appendix

Table: Final training parameters for benchmark CNN-LSTM model.

Parameter	**Final value**
GPU batch size	16
Optimiser	Adam
Loss function	Categorical cross-entropy
Learning rate	1×10^{-5}
Validation approach	Leave-two-out 10-fold cross validation
Evaluation metrics	Accuracy, F1-score

References and further reading

S. J. Kim, K. Koh, S. Boyd, and D. Gorinevsky.

ℓ_1 Trend Filtering.

SIAM Review, vol. 51, no. 2, pp. 339–360, May 2009.

J. Lehtinen, J. Munkberg, J. Hasselgren, et al.

Noise2Noise: Learning Image Restoration Without Clean Data.

ArXiv, Oct. 29, 2018.

M. Irfan, J. Zheng, M. Iqbal, and M. H. Arif.

A Novel Feature Extraction Model to Enhance Underwater Image Classification.

In C. Brito-Loeza, A. Espinosa-Romero, A. Martin-Gonzalez, and A. Safi (Eds.), *Intelligent Computing Systems*, vol. 1187, pp. 78–91.

References and further reading

O. Ronneberger, P. Fischer, and T. Brox.

U-Net: Convolutional Networks for Biomedical Image Segmentation.

May 18, 2015.

M. Irfan, J. Zheng, Jiangbin Zheng, et al.

DeepShip: An Underwater Acoustic Benchmark Dataset and a Separable Convolution-Based Autoencoder for Classification.

Expert Systems with Applications, vol. 183, p. 115270, Nov. 30, 2021.