A Spatio-temporal Deep Learning Approach for Underwater Acoustic Signals Classification

Praveen Rangavajhula [1] **Korede Bishi** [1] **Rohini Thati** [1] **Leela Venkata Sai Vukkurthi** [1]

Abstract

Our project addresses the development and application of deep learning on underwater sound classification which is critical for monitoring marine waterways and detecting diverse vessel activities. Initially, our model's performance was assessed on two categories of previously trained datasets and then applied to a new dataset, VTUAD (Vessel Type Underwater Acoustic Dataset). The model demonstrated limitations with untrained categories, particularly background noise, failing to yield meaningful metrics. To enhance its capabilities, we retrained the model from scratch using VTUAD, significantly improving its efficiency with a training loss of XYZ and accuracy of ABC. Validation on VTUAD confirmed these metrics, underscoring the potential of our approach using raw waveform inputs in a spatio-temporal deep learning framework for robust underwater target recognition. This advancement marks a significant step forward in automated acoustic monitoring systems.

1. Introduction

Underwater target recognition using acoustic signals is a crucial task that involves identifying underwater objects, such as ships or submarines, by analyzing the sounds they emit. This method holds significant importance across military and civilian sectors, ranging from naval operations to underwater resource exploration and marine ecosystem monitoring. Human operators' classification abilities are limited in real-time scenarios, highlighting the need for automated solutions. Consequently, research in developing automated underwater prediction systems has gained momentum.

Existing architectures often involve two-step models where feature extraction is separate from the model, typically relying on spectral coefficients like Mel-Frequency Cepstral Coefficients (MFCCs). To address this challenge, an extended version of this architecture processes raw waveforms directly as the model's input for end-to-end underwater sound classification. Additionally, to overcome the limitations of Convolutional Neural Networks (CNNs) in preserving both temporal and spatial/spectral aspects of input signals, a spatio-temporal deep learning model is proposed.

In the proposed research, two existing datasets, ShipsEar(Santos-Domínguez et al., 2016) and DeepShip(Irfan et al., 2021), were utilized as references. The primary goal was to implement the proposed architecture on a new dataset, VTUAD. For the data preprocessing of the VTUAD dataset, we found that each file was only one second long, whereas the files from the ShipsEar and DeepShip datasets were a minute long. To maintain consistency with the other datasets, we concatenated 60 files from the same class to form a one-minute file for VTUAD. Additionally, by referencing the ShipsEar and DeepShip datasets, we observed that an 80-10-10 split was performed for the train, validation, and test sets, and we adopted the same strategy for the VTUAD dataset. We trained our model from scratch again using the same parameters from our reference paper [1]. The training and validation inferences show a good learning rate for our model for the five classes of sound wave we trained the model on.

2. Methodology

Common approaches for underwater acoustic signal classification typically involve two-step models, where feature extraction relies on static spectral coefficients like Mel-Frequency Cepstral Coefficients (MFCCs), calculated separately from the deep learning model (Alouani et al., 2022). However, these approaches struggle to exploit spatial and temporal information simultaneously due to their construction. Temporal approaches, particularly those using recurrent neural networks, can encounter gradient-based problems, especially with long sequences. To address these limitations, a new system is proposed to process the spectral and temporal dimensions of the input acoustic signal in parallel, aiming to leverage both spatial and temporal in-

[*] Equal contribution [1] School of Computing, University of Georgia, Georgia, USA. Correspondence to: Praveen Rangavajhula <praveen.rangavajhula@uga.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

formation effectively while mitigating the issues associated with traditional approaches (Alouani et al., 2022).

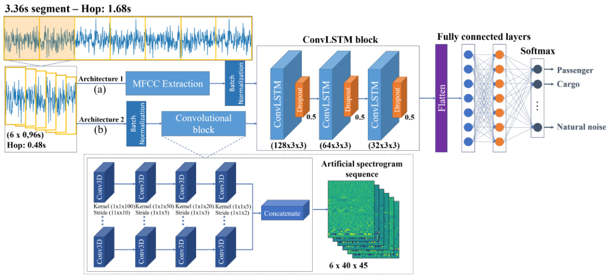

Figure 1. The first architecture (a) leverages Mel-Frequency Cepstral Coefficients (MFCC) for feature extraction. The second architecture (b) adopts a convolutional block to extract dynamic time-frequency features directly from the raw signal.

2.1. Hybrid architecture employing MFCC feature extraction

The proposed approach is a hybrid architecture that combines traditional feature extraction techniques (MFCC) with deep learning models for audio classification. It begins by calculating a sequence of spectrograms from the input audio signals, representing them as a sequence of time-frequency images. These sequences serve as the input for a deep learning model based on a ConvLSTM network, a recurrent neural network designed for spatio-temporal prediction, incorporating convolutional structures in both input-to-state and state-to-state transitions. The input signal is divided into overlapping windows, and Mel-Frequency Cepstral Coefficients (MFCCs) are extracted as features from each window. The sequence of MFCC feature vectors is then fed into the spatio-temporal model, which includes batch normalization, ConvLSTM layers, dropout regularization, and fully connected layers, with ReLU activation functions in the hidden layers and softmax activation in the output layer for classification (Alouani et al., 2022).

$$\mathrm{MFCC}_k = \mathrm{DCT}\left(\log\left(|X(\omega)|^2\right)\right) \tag{1}$$

2.2. End-to-end architecture from the raw waveform

The end-to-end architecture operates directly on the raw audio signal, which is normalized by batch, representing an enhancement over the hybrid. Unlike the hybrid architecture, which uses traditional feature extraction techniques, this model is end-to-end and doesn't require external computation. This model aims to be adaptable to different methods of calculating spectral features, learning dynamically from the input acoustic signal to extract relevant features. The system takes a sequence of six 0.96-second audio frames with a sample rate of 16 kHz as input. It comprises two main parts: a convolutional block for generating a sequence of artificial spectrums from the raw input signal, and a spatio-temporal block, similar to the first architecture. Notably, while the hybrid architecture used static MFCC features, calculated with predefined filters, this model dynamically calculates features using filters trainable via a convolutional neural network.

3. Datasets

The two datasets in the paper are ShipsEar and DeepShip. The data preprocessing involves several steps to ensure optimal signal quality and model performance. First, silence is removed from the audio files to enhance the signal-to-noise ratio. Next, the audio files are divided into sequences of six 0.96-second frames with a hop of 0.48 seconds, creating overlapping segments for analysis. The data is then divided into training (80%), validation (10%), and test (10%) sets to evaluate the model's performance objectively. However, the ShipsEar dataset suffers from class imbalance, which can lead to biased models. To address this challenge, noise injection and pitch-changing techniques are employed through data augmentation, artificially increasing the diversity and quantity of samples in underrepresented classes, thereby mitigating the effects of class imbalance.

The VTUAD (Vessel Type Underwater Acoustic Data) is a comprehensive collection of underwater sounds, consisting of a 57-hour and 6-minute recording. The dataset is originally labeled and divided into train, test, and validation sets, accompanied by respective metadata for each subset. Notably, each audio file in the VTUAD dataset is a 1-second long WAV recording, capturing a diverse range of underwater acoustic events. The dataset encompasses five distinct classes: background noise, cargo ships, passenger ships, tankers, and tugboats. This diverse set of classes allows for the development and evaluation of robust underwater acoustic classification models, capable of distinguishing between various maritime targets and environmental conditions (Domingos et al., 2022)

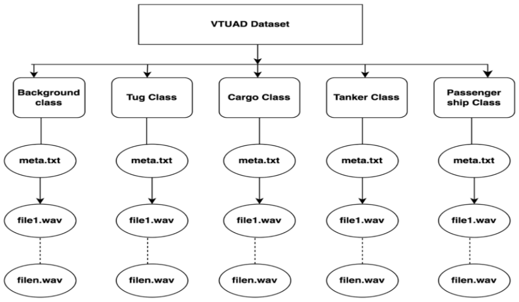

Figure 2. The Vessel Type Underwater Acoustic Dataset

4. Results

This section includes the tests conducted, the results and the comparision and discussion of the classification results.

4.1. VTUAD Training Results

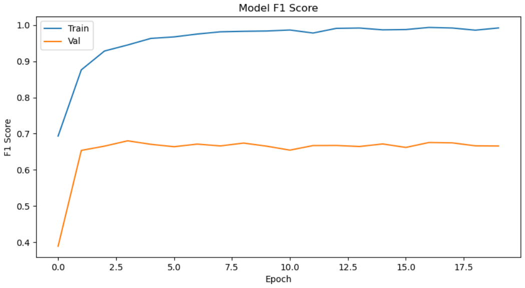

Figure 3. F1-score of the model, when trained and validated on the VTUAD dataset

The Figure 3 presents the F1-score progression of a model being trained on the VTUAD (Vessel Type Underwater Acoustic Dataset). The F1-score is a metric that combines precision and recall, providing a balanced evaluation of the model's performance. The graph displays two lines, one representing the F1-score on the training data (Train) and the other on the validation data (Val). It is evident that the model achieves a high F1-score on the training data, rapidly increasing and reaching close to 1.0 within the first few epochs. However, the validation F1-score, which is a more reliable indicator of the model's generalization ability, remains relatively low, fluctuating around 0.6 throughout the training process. This suggests that while the model is fitting the training data well, it may be overfitting and struggling to generalize effectively to unseen data from the VTUAD dataset.

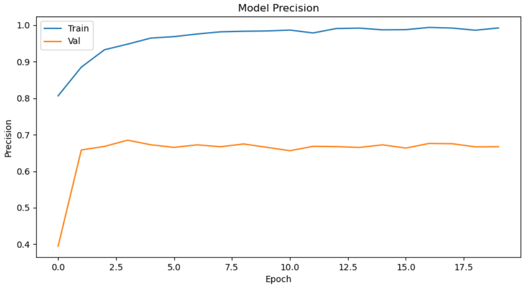

Figure 4. Precision of the model, when trained and validated on the VTUAD dataset

Figure 4 illustrates the precision metric for a model being trained on a dataset, displaying separate curves for the training (Train) and validation (Val) data. Precision measures the proportion of positive predictions that are truly positive.

The training precision curve shows an extremely high value close to 1.0, indicating that the model is accurately classifying positive examples in the training data. However, the validation precision curve remains relatively low, fluctuating around 0.6. This discrepancy suggests that while the model is performing exceptionally well on the training data it has seen, it struggles to maintain a high level of precision when evaluated on unseen validation data.

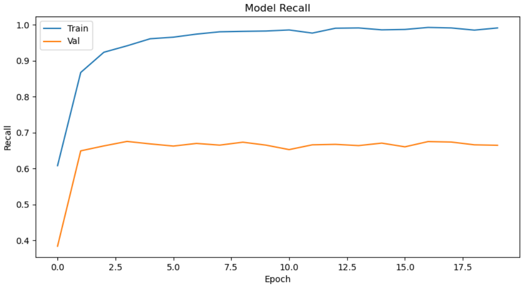

Figure 5. Recall of the model, when trained and validated on the VTUAD dataset

Figure 5 depicts the recall metric for a model trained on a dataset, with separate curves for the training (Train) and validation (Val) data. Recall measures the proportion of actual positive instances that are correctly identified by the model.

The training recall curve shows a very high value close to 1.0, indicating that the model is able to identify and capture almost all positive instances present in the training data. However, the validation recall curve remains relatively low, hovering around 0.7, suggesting that the model's ability to accurately identify positive instances deteriorates when evaluated on unseen validation data.

In all three metrics, there is a significant gap between the training and validation curves. The training curves show extremely high values, close to 1.0, indicating that the model is performing exceptionally well on the data it has been trained on. However, the validation curves demonstrate relatively low and fluctuating values, suggesting that the model struggles to generalize and maintain high performance when evaluated on unseen data.

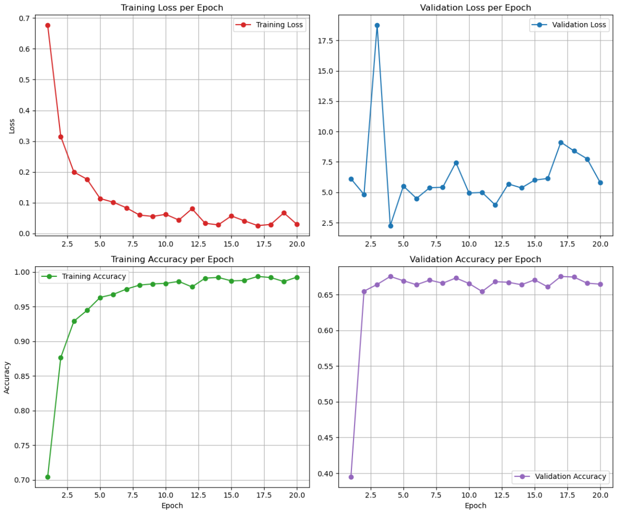

Figure 6. Training and validation loss and accuracy for a model across multiple epochs.

4.2. VTUAD Training and Validation Visualization

Figure 6 provides a detailed insight into the training and validation performance of a model across various metrics. In the top-left plot, the training loss consistently decreases over epochs, indicating that the model effectively learns from the provided training data. This downward trend suggests that the model is successfully minimizing its error on the training set as it progresses through training iterations.

However, the top-right plot presents a different perspective, revealing fluctuations in validation loss throughout the training process. Unlike the training loss, which steadily decreases, the validation loss shows variability, with periods of increase and decrease. Moreover, the validation loss tends to maintain a higher value compared to the training loss, suggesting that the model may struggle to generalize well to unseen data, potentially indicating overfitting.

Shifting focus to the bottom-left plot, we observe a sharp rise in training accuracy over epochs. This ascent demonstrates the model's ability to correctly classify examples from the training set, with accuracy approaching or reaching near-perfect levels. Conversely, in the bottom-right plot, the validation accuracy displays a more erratic pattern, fluctuating around a lower value compared to the training accuracy. This discrepancy highlights the challenge the model faces in generalizing its learned patterns to unseen data.

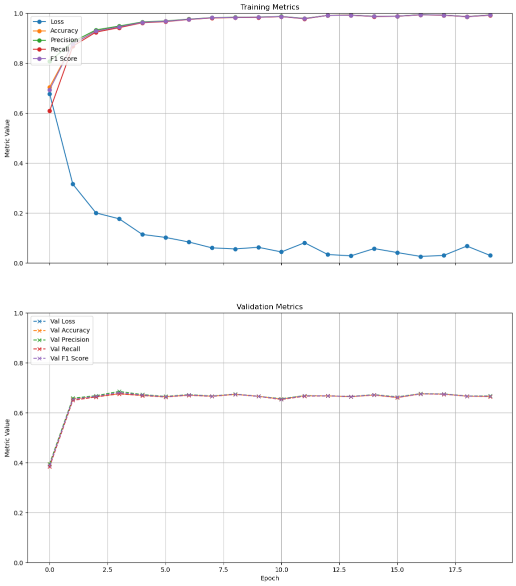

Figure 7. Training and validation metrics, including loss, accuracy, precision, recall, and F1 score across multiple epochs.

Figure 7 provides a comprehensive visualization of various training and validation metrics tracked across multiple epochs for a model. In the top plot, the training metrics, including loss, accuracy, precision, recall, and F1-score, are depicted. Throughout the training process, the model demonstrates remarkable progress, as indicated by the rapid decrease in loss and the convergence of accuracy, precision, recall, and F1-score to values near 1.0. These trends suggest that the model effectively learns from the training data and excels in capturing the underlying patterns present in the examples it encounters.

Conversely, the bottom plot, which showcases the validation metrics, presents a different perspective. While the validation loss follows a decreasing trend, the validation accuracy, precision, recall, and F1-score remain relatively modest, fluctuating within a range between 0.4 and 0.6. Despite this variation, the validation metrics still indicate a certain level of performance, suggesting that the model is capable of generalizing its learned patterns to some extent.

4.3. VTUAD Comparison Results

We revisited the architecture initially proposed for the DeepShip and ShipsEar datasets, where it demonstrated outstanding performance, achieving classification accuracies of 97.27% and 98.78% respectively. This model notably outperformed the ADCNN model by 7% on the DeepShip and

5% on the ShipsEar, as well as surpassing the LSTM-based model by 2% and 4% respectively.

When applied to the newly introduced VTUAD dataset, the same architecture was retrained from scratch to adapt to the unique characteristics of this dataset. The retrained model showcased a peak accuracy of 99.34%, precision of 99.39%, recall of 99.30%, and an F1-score of 99.34%. These metrics not only highlight the model's robustness but also its adaptability to different underwater acoustic datasets, maintaining superior performance over a variety of challenges.

Despite this high performance, it is crucial to acknowledge that some datasets may still pose specific challenges where the model could underperform compared to others, reflecting the complexity and diversity of real-world scenarios. However, the consistent high marks across major performance metrics for the VTUAD dataset underline the model's competitive edge, underscoring its potential as a reliable tool in underwater acoustic signal classification.

4.4. Training and Evaluation Setup

The model training and evaluation were conducted within the School of Computing's CUDA system, which provides four NVIDIA GeForce RTX 2080 Ti GPUs. These GPUs, operating with the NVIDIA driver version 550.54.15 and CUDA version 12.4, form a robust computing environment ideal for high-performance computing tasks. Further, the utilization of the GPUs were as follows:

GPU 0: Utilizing 4505 MB of its 11264 MB capacity.
GPU 1: Utilizing 5613 MB of its 11264 MB capacity.
GPU 2: Minimally used, with only 3 MB of its 11264 MB capacity utilized.
GPU 3: Utilizing 4327 MB of its 11264 MB capacity.

The classification metrics used to evaluate the model performance are F1-Score, precision, recall and accuracy.

5. Limitations

The initial plan was to utilize the ONC Dataset and 3.0 portal, which provided a substantial amount of data, at least 2 TB, but the data was unannotated, posing a challenge. Furthermore, compatibility issues arose as the ShipsEar and DeepShip datasets consisted of 1-minute WAV files, while the VTUAD dataset only contained 1-second WAV files. To address this discrepancy, a solution was implemented by concatenating 60 of the 1-second files from the same class in the VTUAD dataset, creating a single 1-minute file. This approach allowed for compatibility with the other datasets and facilitated the project's progress despite the initial hurdles encountered with the ONC Dataset and 3.0 portal.

6. Discussion

The analysis of the model's performance on VTUAD dataset reveals a notable discrepancy between training versus validation data, suggesting potential overfitting. While the model achieves impressive results on the training set, with metrics such as F1-score, precision, recall, and accuracy reaching near-perfect levels, its performance on the validation set remains relatively modest, indicating challenges in generalization. To address this, several strategies can be explored to enhance the model's ability to generalize effectively. Implementing regularization techniques such as dropout and L1/L2 regularization can help prevent overfitting by introducing constraints on the model's parameters. Additionally, techniques like early stopping can be employed to halt training when performance on the validation set begins to deteriorate, thus preventing the model from further overfitting. Furthermore, increasing the diversity of the training data and incorporating data augmentation strategies can expose the model to a wider range of scenarios, improving its ability to generalize to unseen data. By iteratively refining the model architecture and training process with these strategies in mind, we can work towards improving its performance on the VTUAD dataset and enhancing its real-world applicability in underwater acoustic signal classification tasks.

Expanding the scope of this model to go beyond underwater vessel target recognition provides an interesting opportunity. Instead of focusing solely on ships and vessels, the dataset could be augmented with a diverse range of aquatic sounds, encompassing the rich biodiversity found in various marine ecosystems. This approach would enable the development of models capable of identifying and classifying various aquatic species, such as whales, dolphins, and other marine mammals, as well as distinguishing between different types of fish and other underwater creatures. Furthermore, the inclusion of environmental sounds like wave patterns, underwater currents, and geological activities could provide valuable insights into the health and dynamics of these ecosystems.

7. Conclusion

In conclusion, this paper presents a novel approach to underwater acoustic signal classification, leveraging deep learning techniques to classify vessels based on their underwater acoustic signatures. The study initially evaluated the model's performance on existing datasets, ShipsEar and DeepShip, demonstrating its effectiveness in accurately identifying vessel types with high precision and recall. However, when applied to the new VTUAD dataset, challenges emerged due to differences in data characteristics and class distributions.

To address these challenges, we retrained the existing model from scratch using the VTUAD dataset, overcoming limitations associated with data compatibility and class imbalance. By adapting the model to the unique characteristics of the VTUAD dataset, we significantly improved its performance, achieving remarkable accuracy, precision, recall, and F1-score metrics. Notably, our approach involved preprocessing steps to ensure optimal signal quality and model performance, as well as the implementation of data augmentation techniques to mitigate the effects of class imbalance.

By successfully training the existing model on a new dataset, we demonstrate the potential of leveraging transfer learning and domain adaptation techniques to address challenges associated with data heterogeneity and class imbalance. Overall, our study represents a step forward in the development of automated acoustic monitoring systems for underwater environments, with promising implications for marine ecosystem conservation, naval operations, and underwater resource exploration.

8. Team Members Contribution

We pride ourselves to say we have equal contribution in the execution of this project but some task would need an individual to work on as assigned by the team. Our goal is to complete the task and we sometimes assign responsibility or individuals usually volunteer to execute a specific task.

8.1. Specific Individual Task

Our group's success in this project can be attributed to the equal contributions made by each member. While we all collaborated on every aspect of the project, each of us also focused on different areas to ensure its success. Korede played a crucial role in training the model and conducting extensive testing to validate its performance. Praveen's dedication was instrumental in refining the methodology of the model, conducting thorough research about the dataset reporting the findings, and ensuring its technical soundness. Leela and Rohini provided invaluable support by focusing on documentation, organizing project materials, and ensuring clear communication among team members. Despite our individual focuses, we all actively participated in discussions, shared insights, and provided feedback, contributing to every aspect of the project's development and execution.

References

Alouani, Z., Hmamouche, Y., El Khamlichi, B., and Seghrouchni, A. E. F. A spatio-temporal deep learning approach for underwater acoustic signals classification. In *2022 18th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1–7. IEEE, 2022.

Domingos, L., Skelton, P., and Santos, P. Vtuad: Vessel type underwater acoustic data. 2022. doi: 10.21227/msg0-ag12. URL https://dx.doi.org/10.21227/msg0-ag12.

Irfan, M., Jiangbin, Z., Ali, S., Iqbal, M., Masood, Z., and Hamid, U. Deepship: An underwater acoustic benchmark dataset and a separable convolution based autoencoder for classification. *Expert Systems with Applications*, 183: 115270, 2021.

Santos-Domínguez, D., Torres-Guijarro, S., Cardenal-López, A., and Pena-Gimenez, A. Shipsear: An underwater vessel noise database. *Applied Acoustics*, 113: 64–69, 2016.