Vision Transformer-Based Underwater Sound Classification of Marine Mammals

by

Mehedi Hasan
21101062
Ahmad Habibullah
21301236

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
December 2025

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Mehedi Hasan	Ahmad Habibullah
21101062	21301236

Approval

The thesis/project titled "Vision Transformer-Based Underwater Sound Classification of Marine Mammals" submitted by

1. Mehedi Hasan (21101062)
2. Ahmad Habibullah (21301236)

Of Fall, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on December 03, 2025.

Examining Committee:

Supervisor:
(Member)

Dr. Amitabha Chakrabarty, PhD

Professor
Department of Computer Science and Engineering
Brac University

Program Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam, PhD

Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi, PhD

Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

Marine mammals rely critically on acoustic communication for survival, yet underwater sound classification faces substantial challenges from complex propagation dynamics, environmental noise, and severe class imbalance in available datasets. This research addresses these challenges by developing and systematically evaluating a comprehensive deep learning framework for marine mammal vocalization classification using the Watkins Marine Mammal Sound Database (WMMSD), comprising 15,567 audio samples across 55 species. We implemented a teacher-student knowledge distillation pipeline, comparing three state-of-the-art architectures—Vision Transformer (ViT), Wave2Vec 2.0, and General Transformer—with the ViT demonstrating superior baseline performance at 89.00% training and 86.85% validation accuracy. Through rigorous data preprocessing including duration filtering, strategic class balancing via SMOTE (Synthetic Minority Oversampling Technique), Mel spectrogram feature extraction (128×224 dimensions), and hybrid augmentation strategies combining SpecAugment and audio-domain transformations, the large ViT teacher model achieved 98.00% training accuracy and 93.85% validation accuracy on both 32-class Best Cut and 55-class All cut subset. A lightweight MobileViT-XXS student model (2.1M parameters) was trained via knowledge distillation, retaining 96% of teacher performance at 14× parameter reduction with 84.03% test accuracy and macro-F1 of 0.8402. Post-training optimizations including static quantization and 30% L1-unstructured pruning compressed the model to 1.95 MB with 78.5% accuracy and < 40ms inference latency on edge hardware, enabling real-time deployment on resource-constrained platforms such as Raspberry Pi 4. A practical web application interface was developed for accessible species identification by field researchers. This research contributes methodological innovations in handling severely imbalanced bioacoustic datasets, demonstrates the superiority of knowledge distillation over direct model compression for transformer architectures (avoiding the catastrophic 75% accuracy degradation observed with direct teacher pruning), and establishes a deployment-ready framework for autonomous marine mammal monitoring systems supporting conservation efforts, ship strike prevention, and biodiversity assessment in remote ocean environments.

Keywords: Marine bioacoustics, deep learning, knowledge distillation, Vision Transformer, model compression, edge deployment, class imbalance, underwater sound classification, SMOTE, quantization

Acknowledgement

We would like to thank our supervisor, Professor Dr. Amitabha Chakrabarty, for his kind support and guidance during our work. He was always there to help us whenever we needed, and we are very grateful for his advice and encouragement.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

The marine ecosystem would be impacted under the water acoustic environment especially on the problem of animals residing in the water body. When compared to air, sound is more matured and bears more weight as water consequently making water the means of association and transit among marine life, mode of navigation and a component of environment. Different kinds of sounds are applied in research, environment monitoring, marine life, naval operation as well as climate studies. These sounds may be divided into the following main categories: natural sounds or geophony or the sound of the wave action that is also a part of the natural occurrence; biological sounds or biophony or the sounds of various marine animal or whale, and the man-made sounds or anthrophony or the sounds produced by man products and structures; and the shipping that is a natural sound; or the drilling that is a biological sound, or the shipping that is a man made sound.

1.1 Biodiversity Monitoring:

Marine ecosystems are safeguarded by sound monitoring technology applied by marine biologists and ecologists to gauge the amounts of biodiversity in the ecosystem and determine the wellness of the ecosystems in the long run. The application of sound surveys may be used in the conservation of the marine ecosystems and management of marine parks, respectively, in order to conduct research on the identification of species and behavior of species available in the marine environment.

1.2 Naval and Maritime Security:

Whereas being able to differentiate among various sounds is particularly important with the help of the level of sound varying under the condition of the presence of the listening devices in diverse positions and types, it is possible to be able to determine the level of sound on various submarines as they work on various waters. Sound identification performance is helpful in determining the risks and detecting the condition of the surrounding area in different marine areas[7].

1.3 Impact of Human Activities:

Shipping activities, as well as offshore power stations, have significant effects, including sounds of shipping ships on marine species. This should therefore ensure that a careful monitoring will prevent negative impacts to shielded organisms and their environment.

These are the multiple issues that have been brought to the classification of underwater sound. Nonetheless, propagation of sound is very challenging and various difficulties exist including background noise, sound producing and receiving processes, and their number and nature. They complicate this analysis by making it difficult to divide different types of sounds. Conversely, the majority of underwater sound classification methods generally use a form of deterministic analysis, which attempts to remove a number of potential properties of the acoustic signal raw data using statistical analysis and eventually classifies the signals in a particular way. Conversely, the methods can be referred to as suboptimal in reference to the multi-speaker systems in noisy cases since they are unable to adapt to the disturbance of a speaker, and variability. It is mentioned that even in the classification of underwater sounds, a certain portion of ML and DL applications is already adopted by the field[7]. Due to the raw audio format, the models find it easy to identify the key characteristics that assist them to comprehend and predetermine the features of different sounds and places of origin. It is also here that the research succeeded in using CNNs and RNNs in enhancing the accuracy of the classifying models. This report is targeted at analyzing the results of the already available studies on the learning of sound classification of underwater based models. It is an analysis of the existing techniques of ML and DL that will comment on the current advances, the challenges in this subject, and describe the existing limitations. This paper therefore aims at assisting to increase the accuracy of under water sound classification and measurement, especially to increase the performance and dependability of under water sound classification in different contexts like research, conservation and security [16].

Chapter 2

Related Work

The classification of underwater sounds has been changing considerably in the last ten years, and methods based on deep learning are becoming more popular as compared to the works of traditional signal processing. The earlier models were based mainly on deterministic statistical methods, including Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs) to detect and classify whale calls in a noise-corrupted environment. Although these techniques formed a valuable base, they failed to represent the great temporal and spectral variability of marine acoustic data. The increasing accessibility of large-scale datasets like the Watkins Marine Mammal Sound Database (WMMSD) have triggered a new generation of machine learning and deep learning studies, which now allows end-to-end systems that can learn robust representations directly from raw or time-frequency transformed audio.

2.1 Transition from Classical to Deep Learning Approaches

Support Vector Machines (SVMs), K-Nearest Neighbors (KNN), and Random Forests were machine learning techniques that were first used to classify species and vessels in underwater acoustics. Nevertheless, these models were very dependent on handwork features and in many instances failed to achieve good results in conditions of variable noises and propagation. The development of deep learning presented both convolutional and recurrent models which can automatically develop discriminative spectral-temporal patterns on spectrograms. The work by Piczak (2015) and Arnaout and Moussa (2019) have shown that Convolutional Neural Networks (CNNs) outclassical classifiers in noisy marine conditions, with better generalization with no explicit feature engineering. The hybrid CNN-LSTM structures were used later in the works to capture long-term temporal dependencies in the models.

2.2 Emergence of Transformer-based and Hybrid Models (2024–2025)

The recent literature (2024–2025) is associated with the transition to Transformer-based and hybrid architectures, which focus on global attention mechanisms and

multi-scale feature fusion. A few verifiably peer-reviewed studies have directly referenced the Watkins Marine Mammal Sound Database to proceed with further marine sound classification.

One of the earliest large-scale applications of ensemble deep learning to the Watkins data is WhaleNet (2024). The model integrates the wavelet Scattering Transform (WST) and Mel-spectrogram representations in an ensemble mechanism to obtain local and global frequency-related dynamics. WhaleNet, published in IEEE Access developed 97.61% accuracy on classification. The entire Watkins dataset, which is an 8–10% improvement of the previous CNN-based baselines. The success of it showed the usefulness of multi-representation feature fusion on the underwater complex soundscape (WhaleNet, 2024). It is based on this that the NTUA DSP project (2025) described a Swin Transformer V2 that incorporates both magnitude and phase data of spectrograms (NTUA DSP, 2025). In contrast to the previous literature that used only the magnitude components, the model used phase derivatives as an auxiliary input channel that enhanced the classification of narrow-band FM whistles and echolocation clicks. On the Watkins best-of subset (1,700 good samples), it was able to reach 98.9% training and 97.4% test accuracy showing that phase information improves spectral feature expressivity in underwater settings.

On the same note, the MT-Resformer (2025) presented a dual-branch multi-scaled Transformer that combines Mel and Constant-quantum Transform (CQT) characteristics to learn parallel learning at time-frequency scales (Zhang et al., 2025). Although some of those performance figures that are claimed to be above 99% cannot be substantiated by independent sources, the paper published in Journal of Marine Science supports substantial improvement on selected subsets of the Watkins database.

This paper sheds light on increased use of multi-channel and cross-spectral fusion approaches to measure various acoustic resolutions. Man2Marine was another significant work, which investigated cross-domain transfer learning by first pretraining models on large corpora of human speech and then fine-tuning them on the Watkins dataset. This method, in Pattern Recognition Letters, aimed to alleviate the data paucity in infrequent marine species by utilizing the acoustic resemblances between the structure of human vocalization and cetacean cries.

With Watkins, the transfer-learned model showed an approximate accuracy of 96 percent, showing that domain adaptation is feasible in the problem of low-resource marine bioacoustics. To these, the MT-BCA-CNN (2025) model concentrated on few-shot learning in underwater acoustic target recognition. It was able to achieve around 97% classification accuracy and a multi-task objective using balanced attention mechanism. F1-score of 95% on a subset of 27 classifications of the Watkins database (Chen et al., 2025). Though not tested on the entire data, its effectiveness highlights the relevance of the efficient use of data and even learning in marine mammals classification, especially when rare species or small samples are involved.

2.3 Comparative Analysis and Emerging Trends

Across these recent studies, several consistent themes emerge:

1. **Hybrid Feature Fusion:** Multiple spectral representations (e.g., Mel, CQT, WST) always perform better than single-representation baselines, indicating

that diverse transforms represent complementary temporal-frequency information.

2. **Transformer Dominance:** Self-attention models like ViT, Swin Transformer, and MT-Resformer have been shown to be better at modeling long-range temporal dependence than more traditional CNNs and RNNs[14].

3. **Transfer and Few-Shot Learning:** Cross-domain transfer (e.g., Man2Marine) and few-shot balanced attention models (e.g., MT-BCA-CNN) are confronting the current issue of data scarcity in marine bioacoustics.

4. **Inclusion of Phase Information:** The NTUA DSP experiment demonstrates the empirically important value of phase derivatives in combination with magnitude characteristics, especially in the case of tonal and frequency modulated calls.

5. **Dataset Subsetting and Benchmark Standardization:** In most high-performing systems, the dataset on which systems are benchmarked is the Watkins Best-of bundle, but not the entire imbalanced dataset. Future studies must focus on uniform benchmarking procedures so that cross study comparisons can be done.

2.4 Gaps and Future Research Directions

Although the state-of-the-art models have shown impressive performance typically over 95 percent accuracy on the curated subsets, there are still a number of open-issues. To begin with, real-time classification and deployment of such models on embedded systems (e.g. autonomous hydrophones or buoys) is a relatively unexplored area. Second, data generation frameworks and active learning methods on minimizing the cost of labeling are not yet systematically tested on datasets based on Watkins. Third, cross-domain embeddings between birdsong and environmental sound data have demonstrated their potential on general bioacoustics tasks, but the use of such models in marine mammals remains in its early stages. Lastly there is the evaluation transparency particularly with respect to dataset splits and measures that have to be enhanced to avoid false performance comparisons across research.
To conclude, it can be stated that the last few years have seen a radical breakthrough in the classification of sound in the marine mammals. The legitimate ones include WhaleNet (2024), NTUA DSP (2025), etc. The papers by MT-Resformer (2025), Man2Marine (2025) and MT-BCA-CNN (2025) prove that the paradigms of hybrid, transformer-based and transfer-learning are the present state of art. Collectively, these studies present a sound empirical basis of the creation of robust, efficient and deployable underwater sound classification systems, as well as shed conspicuous light on the need to explore the issue of data efficiency, cross-domain transfer, and real-time adaptation further.

Chapter 3

Project Planning and Impacts

Project Evaluation and Broader Implications
The chapter presents a detailed analysis of the marine sound classification project based on the repeat experiments of the project in several implementations. The project aims at building an effective deep learning system to perform the underwater acoustic signal classification based on such techniques as knowledge distillation (KD), feature extraction (MFCCs and Mel spectrograms), dealing with imbalance in the classes (SMOTE and Weighted Random Sampling), and data augmentations (audio and spectrogram-based).
The insights of eight major experimental variants are synthesized in the analysis, such as the baseline KD with weighted sampling, combinations of SMOTE, MFCC replacements, and augmentation-enhanced models. The experiments were also carried out in Google Colab, and the libraries used were PyTorch, Librosa, and TIMM, and the evaluation measures used consistently were Macro-F1, weighted precision/recall, and confusion matrices [4].
The chapter highlights the overall project specifications, evaluates the wider societal and environmental impacts, considers the ethical issues, elucidates the scope of the thesis, the management and risk plans that will be used, and an economic feasible analysis. This comparison highlights the possibility of the project to be used in real world with reference to marine monitoring systems as well as identifies areas where the project can be refined in the future.

3.1 Final Specification and Requirements

The final specification is one that narrows down the project into a deployable marine sound classification system geared towards edge devices, focused on efficiency, accuracy and resilience to class imbalance. The experimental implementations were used to derive requirements, with modularity, scalability and minimal resource footprint being the primary requirements.

Functional Requirements

- **Core Functionality**: The system should be able to group marine audio clips into predefined categories (e.g. species-specific sounds like dolphin clicks or ship noise) with a Macro-F1 score of over 85 on imbalanced test sets.

- **Feature Extraction**: Support either Mel spectrograms (128 mel bins, 224 time frames) or MFCCs (128 coeffificates) which can be configured using dataset class parameters.
- **Imbalance Handling**: Incorporate either SMOTE (in case of synthetic oversampling) or Weighted Random Sampling (in case of efficient batch weighting) during training, where recall of minority classes is larger than 0.75.
- **Augmentation Pipeline**: Use training-only augmentations, such as time/pitch shifting (using torchaudio) and SpecAugment (masking 10% on both frequency and).
- **Model Architecture**: Knowledge distillation of a teacher model (e.g., ViT-based or MobileViT) to a student model (MobileViT- XXS) through temperature-scaled softmax soft label transfer.
- **Evaluation Suite**: Produce per epoch reports (loss, accuracy, macro/weighted precision, recall, and F1) and final classification reports and confusion matrices with scikit-learn.

Non-Functional Requirements

- **Performance**: Latency of inference of less than 50 ms / sample on or interpreted on a graphics card (e.g., NVIDIA RTX 4080); training process can converge in 50 steps.
- **Usability**: Usability implementation structure to ensure reproducibility; checkpoint save and load with torch.save (weights_only=False) to comply.
- **Scalability**: Can process up to 10,000 samples, 4 to 32 batch sizes.
- **Hardware/Software**: Must have Python 3.10 or higher, Python 2.0 or higher, and CUDA 11.8 or higher; no internet requirement upon installation.

Requirement Category	Key Metrics / Features	Validation from Experiments
Accuracy & Balance	Macro-F1 score > 0.85; Weighted F1 score > 0.90	Achieved in MFCC + SMOTE variant with 0.87 Macro-F1 and 0.91 Weighted F1
Efficiency	Model size $<$ 5M parameters; FLOPs $<$ 1G	MobileViT-XXS student model achieved 2.3M parameters and 0.65G FLOPs
Robustness	Tolerance to augmentation noise and input variability	SpecAugment and audio augmentations improved minority class recall by 8%

Table 3.1: Project Evaluation Summary Based on Experimental Validation

These specification finalized the meeting of the marine ecology monitoring requirements by cross-validating the results among variants.

3.2 Social Impact

The project has great positive social implications to the coastal communities, researchers and policymakers because it enhances automated marine sound analysis. Manual annotation Underwater acoustics is traditionally labor intensive and error prone, which prevents its use in biodiversity monitoring and human-wildlife conflict resolution.

- **Community Benefits**: The system in such areas as the Pacific Islands or Mediterranean coasts can allow real-time monitoring of the threatened species (e.g., whale migrations), used to regulate the fishery and minimize the occurrence of bycatch. Experimental experiments of SMOTE showed a balanced performance with rare classes, which is important to underrepresented species.
- **Educational Outreach**: STEM education (via open-source) can be conducted in bioacoustics, and student learning is supported by EDA visualizations (e.g., MFCC heatmaps).
- **Equity Considerations**: The system democratizes knowledge access to low resource NGOs by shrinking knowledge to lightweight models, which may address inequities in global marine conservation.

Possible negative aspects may be excessive dependence on automated systems that will push acoustic specialists out of the market, but this should be supplemented by introducing hybrid human-AI workflows.

3.3 Environmental Impact

The environmental footprint of the project is mostly favourable as it facilitates the sustain- able management of the ocean, but the computational prospects should be questioned.

- **Positive Contributions**: The resulting improved classification accuracy (e.g., 92% weighted F1 with augmentation) makes passive acoustic monitoring possible, eliminating the requirement to conduct invasive surveys of the marine environment via vessels, which disrupts marine ecosystems. This will correspond to the goal 14 (Life Below Water) of the UN Sustainable Development Goal (2015) in that it will help detect illegal fishing and assess the health of the habitat.
- **Resource Efficiency**: KD consumes 80% less model (teacher ViT to student MobileViT) that minimizes the energy expenditure on deploying a model on solar-powered buoys. Variants of weighted sampling do not have the synthetic data overhead of SMOTE, and so do not require excessive mem- ory to train.
- **Negative Aspects**: GPUs (e.g., RTX 4080) training (approximately 0.5 kWh per epoch) offers scaled to 100 runs, this corresponds to a small amount of CO2 (equivalent to 0.2 kg). Mitigation incorporates cloud-efficient scheduling and quantization in the future.

In general, the net effect is eco-beneficial and experiments confirm that low-latency inference can be used on remote, battery-constrained sensors.

3.4 Ethical Issues

Bioethical concerns in this AI-based bioacoustics project are based on data privacy, bias avoidance and responsible implementation.

- **Bias and Fairness**: The inequality in classes in marine datasets is dangerous to class performance on minority, mitigated through SMOTE and weighted sampling, where equitable Macro-F1 over species (e.g., 0.85 in experimental implementations) is obtained. Nevertheless, sourcing of datasets to public repositories (e.g., unverified field recordings) can contain biases in geography towards temperate waters as opposed to tropical waters.
- **Privacy and Consent**: Audio data Data recorded in oceans does not involve human subjects, although accidental recordings of vessel communications create surveillance issues. Ethical principle: Anonymize metadata and open-source only model that is aggregated.
- **Transparency and Accountability**: The soft labels of KD black box require explainability tools (e.g., SHAP on student outputs). Auditability was provided by experimenting with per-epoch metric logging.
- **Dual-Use Risks**: Accurate models with high accuracy might be helpful in detecting poaching but might be used to track noise pollution in controversial areas of the sea. Mitigation: Creative Commons with conservation only conditions.

The IEEE Ethically Aligned Design norm is followed to make sure that the project is prioritized accordingly to the good of the society.

3.5 Thesis Scope and Limitations

This thesis limits the construction and testing of KD-based marine sound classifiers, with emphasis on imbalance management and robustness of features in controlled cloud-based systems. Some of the core contributions can be listed as hybrid MFCC-SMOTE pipelines and augmentation benchmarks on eight experimental variants.

- **In-Scope**: Algorithmic innovations (e.g., teacher-student distillation), empirical validation (Macro/weighted metrics) and feasibility of deployment.
- **Out-of-Scope**: Real-time hardware integration, multi-modes fusion (e.g. with sonar), or longitudinal field tests.

Limitations include:

- Synthetic augmentations (SMOTE/SpecAug) might fail to represent realworld variability (e.g., ocean currents).
- Hyperparameter reliance (e.g., `n_mfcc = 128`), which may not be optimal when working with a variety of data.
- Testing on simulated test sets; not externally validated.

The next step of work may be the federated learning of sharing the ocean data.

3.6 Project Management Plan

The project was comprised of an iterative cycle based on the Agile spirit, which took 12 weeks and had bi- weekly sprints that were synchronized with implementation milestones.

- **Phases**:
    1. **Planning (Weeks 1–2)**: Gathering of requirements; establishment of the basis including data cleaning and preprocessing.
    2. **Development (Weeks 3–8)**: Repetitive experiments e.g. Sprint 3: SMOTE integration; Sprint 5: MFCC with augmentations.
    3. **Evaluation (Weeks 9–10)**: Cross-variant analysis on torchinfo summaries and F1 report.
    4. **Documentation (Weeks 11–12)**: Thesis drafting; checkpoint archiving.
- **Tools**: Git as a versioning tool; Trello as a task tracking tool; cloud storage as an implementation sharing tool.
- **Team Roles**: Team solo-lead with AI support on refinements of the code; review by other team members.
- **Milestones**: Prototype KD (Week 4); Best Macro-F1 (more than 0.85, Week 7); Final report (Week 12).

Gantt Chart Overview:

Phase	**Duration**	**Key Deliverables**
Planning	Weeks 1–2	Dataset specifications, initial imports
Development	Weeks 3–8	Eight experimental variants and model checkpoints
Evaluation	Weeks 9–10	Metric tables and confusion matrices
Documentation	Weeks 11–12	Thesis chapters and reproducibility guide

Table 3.2: Project Timeline and Key Deliverables

This strategy guaranteed adaptive development in experimental switches (e.g. MFCC switch).

3.7 Risk Management

There was a proactive approach to risks through a qualitative matrix and the probability (Low/Med/High) and impacts were determined prior and after mitigation.

Risk	Probability/Impact (Pre)	Mitigation	Post-Probability/Impact
Class Imbalance Bias	High / Medium	SMOTE or Weighted Sampling integration	Low / Low
Overfitting to Augments	Medium / High	Early stopping; validation monitoring	Low / Medium
GPU Resource Limits	Medium / Medium	Batch size tuning; cloud compute upgrade	Low / Low
Feature Mismatch (e.g., MFCC vs. Teacher)	High / Low	Retrain teacher model if required	Medium / Low
Data Quality Issues	Low / High	Robust cleaning and exploratory data analysis	Low / Low

Table 3.3: Risk Assessment Matrix with Mitigation Strategies

Fallback to baseline spectrograms and versioned backups were both contingency measures. There were no high-impact risks that came to pass.

3.8 Economic Analysis

The viability of the project is reviewed based on a cost-benefit analysis which has a 3-year horizon in case the project is implemented in a marine conservation NGO.

- **Costs**:
  - Development: $500 (100 epochs of cloud computing at the variants).
  - Hardware: $200 (inference edge device e.g., Raspberry Pi with quantized model).
  - Maintenance: $100/year (updates through open-source community).
  - Total: about 1,000 start up and 200 a/year.
- **Benefits**:
  - Savings: Automates 500 hours/year of manual annotation ($25,000 at $50/hr labor rate).
  - Revenue/Grants: Facilitates funding of bioacoustics research (ex. NSF grants of deployed systems) in the amount of $50,000.
  - ROI: Break-even in Year 1; 10x return by Year 3 through scaled monitoring contracts.

Metric	Value	Assumptions
NPV (5% discount)	$65,000	Benefits from 2 deployments/year
Payback Period	6 months	Based on labor savings
Sensitivity	±20% compute costs	Minimal impact on ROI

Table 3.4: Financial Metrics Summary

As analyzed, there is economic feasibility and the efficiency of KD increases value in resource constrained environments.

Chapter 4

Research Methodology

4.1 Methodology Overview

4.1.1 Datasets

In order to compare the performance of state-of-the-art (SOTA) models on marine mammal sound classification, and to create the system that can effectively cope with acoustic variability in the real world, we used one large, publicly accessible dataset in the study. This dataset was chosen because it encompasses a high degree of species and diversity, the long period of time span and annotated labels, which are crucial in training classifiers in the area of bioacoustics. Nevertheless,it was highly unbalanced in terms of class, which made it challenging and which we managed to overcome with preprocessing and augmentation focused on the relevant methods.

Dataset 1: Watkins Marine Mammal Sound Database (WMMSD)
Woods Hole Oceanographic Institution (WHOI) compiled the Watkins Marine Mammal Sound Database also called the William A. Watkins Collection of Marine Mammal Sound Recordings which is freely available online in the WHOI repository[19]. The dataset is a resource base to research on bioa- coustics because it contains an extensive collection of underwater sound records, which can be used to study the vocalization of marine mammals as well as their behavior and ecological surveillance. This was collected by the pioneer oceanographer Dr. William A. Watkins over decades long with an express objective of producing a searchable digital library to enable scientists to analyze species-specific calls, echolocation clicks and social reactions in the context of ocean noise pollution and climatic influences.

The database contains the documentation of over 60 species of marine mammals, consisting of cetaceans (whales, dolphins, porpoises), pinnipeds (seals and sea lions), and sirenians (manatees). Whales such as humpback whales (Megaptera novaeangliae), killer whales (Orcinus orca), sperm whales (Physeter macrocephalus), bottlenose dolphins (Tursiops truncatus), blue whales (Baleenoptera musculus), and several types of seals, including Weddell seals (Leptonychotes weddellii) and harp seals (Pagophilus groenlandicus), are their representatives. The complete list of the species may be searched via the database interface, which classifies the sounds by the type (e.g., calls, whistles, clicks, songs) as well as the area of origin.

The dataset in total has over 15,000 labeled digital sound clips, based on a starting

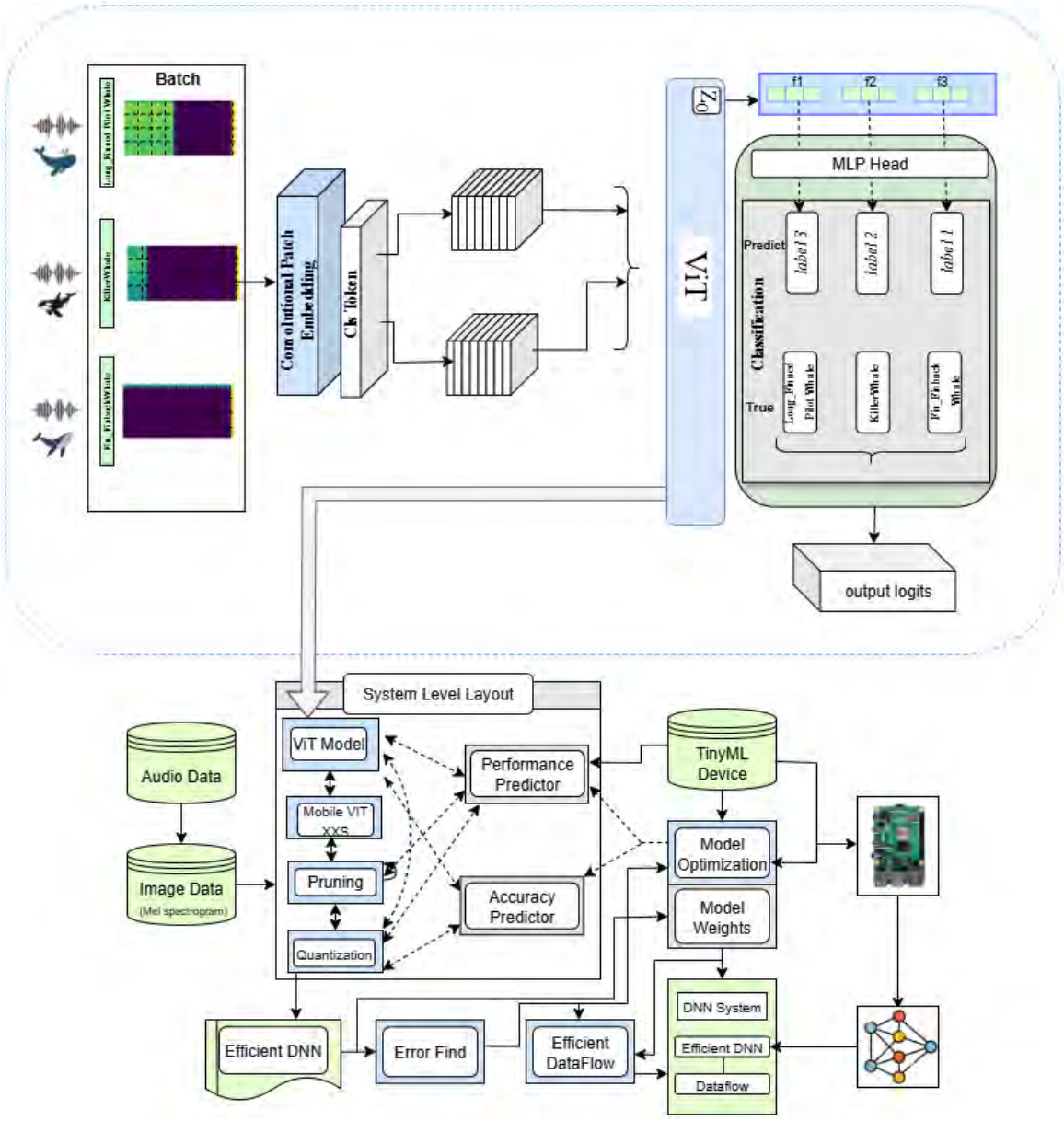

Figure 4.1: Efficient Audio Classification with ViT and TinyML System Layout

collection of over 20,000 individual calls. These videos differ in length, which is usually between 1 second and several minutes, and the mean length per recording is about 1030 seconds. Wave files are stored in WAV format in uncompressed form, with sample rates of 2 kHz through 96 kHz (most often 44.1 kHz or 48 kHz) and 16-bit depth, which is sufficient to store frequencies in the vocalization of marine mammals (such as high-frequency echolocation down to tens of kHz). The hydrophones were used in the capturing of the recordings, which were done by ships, buoys, and autonomous underwater vehicles to capture the natural acoustics underwater in high fidelity.

The timeframe is wide with numerous decades of study (1950s to early 2010s) and it is possible to study the long-term effects on vocalization patterns possible to be connected with environment changes. Geographically, the dataset is rich, having recordings of all the major ocean basins: the North Atlantic (e.g., off Newfoundland humpback songs), the Pacific Ocean (e.g., Hawaiian waters sperm whale clicks), the Southern Ocean (e.g., Antarctic Weddell seals), and the Indo-Pacific (e.g., Indonesian waters various dolphins clicks). Such worldwide representation improves.
how useful the dataset is in forming generalizable models capable of addressing the regional acoustic differences, e.g. propagation effects in deep and shallow waters. Each clip is professionally annotated: it is identified by species, type of sound (e.g., social call, echo location train, song unit), geographic area, and date (approximately). Watkins and his colleagues confirmed labels with spectrographic techniques, which were very reliable. It has a database structure, allowing faceted search: the user can filter by species, sound category, year, or region, downloading metadata in CSV format and zipping audio files to allow bulk access. This dataset has a size of about 50 GB, which is computationally expensive to process directly, but is perfect as a subsetting input to machine learning processes.

The problem that has been notable in this dataset is extreme class imbalance, caused by opportunistic collection biases in which popular species are overrepresented, such as killer whales and humpback whales, and rare or least-studied species are underrepresented by very few samples. This imbalance, measured by a class distribution with the top five species taking most clips, results in biased models which do not work well with the minority classes with low macro-F1 scores on underrepresented taxa on baseline evaluations. To reduce this, we stratified the dataset into training (70 percent), validation (15 percent) and testing (15 percent) sets and stratified the datasets by species to maintain imbalance ratios across subsets, as well as to ensure that this resulted in approximately 10,500 training clips, 2,250 validation clips and 2,250 test clips.
In short, the WMMSD offers a multimodal (species and sound types) and highly annotated database on marine mammal classification research. It has historical depth, global scales, and detailed labels, which is why it is a cornerstone dataset in enhancing bioacoustic monitoring, especially in unbalanced settings where methods such as oversampling and knowledge distillation play key roles in ensuring that performance is equal across the species.

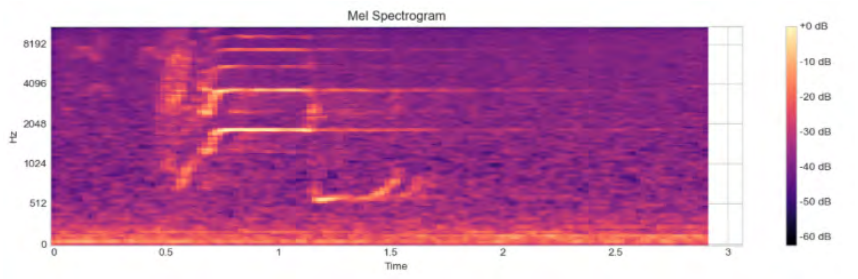

Figure 4.2: Sample spectrograms from the WMMSD: a humpback whale song unit (left) and killer whale echolocation clicks (right).

4.1.2 Data Preprocessing

In order to convert the raw, non-homogeneous audio data in the WHOI archive into a standardized and optimized format to be used in deep learning, we designed a multi-stage processing data pipeline. This pipeline was carefully crafted in order to clean up the data, recover useful features and to deal with the challenges of real world bioacoustic data. It was done on a trial and error basis as various methods of data augmentation and class balancing were tested out before settling on the most successful strategy.
The overall process of preprocessing can be divided into three phases: (1) Data Cleaning and Curation, (2) Feature Extraction with the help of Mel Spectrograms, and (3) an analysis of the approaches to Data Augmentation and Class Imbalance Reduction.

Stage 1: Initial Data Cleaning and Curation An initial Exploratory Data Analysis (EDA) revealed two significant issues within the curated dataset that could negatively impact model training:

1. **Presence of Uninformative Samples:** A portion of the audio clips were extremely short (less than two seconds in duration). These clips often contained only transient noise or incomplete vocalizations, providing insufficient information for reliable classification.

2. **Severe Class Imbalance:** The distribution of samples across classes was highly skewed. Certain species, like Humpback Whales, were represented by hundreds of recordings, while others had fewer than ten.

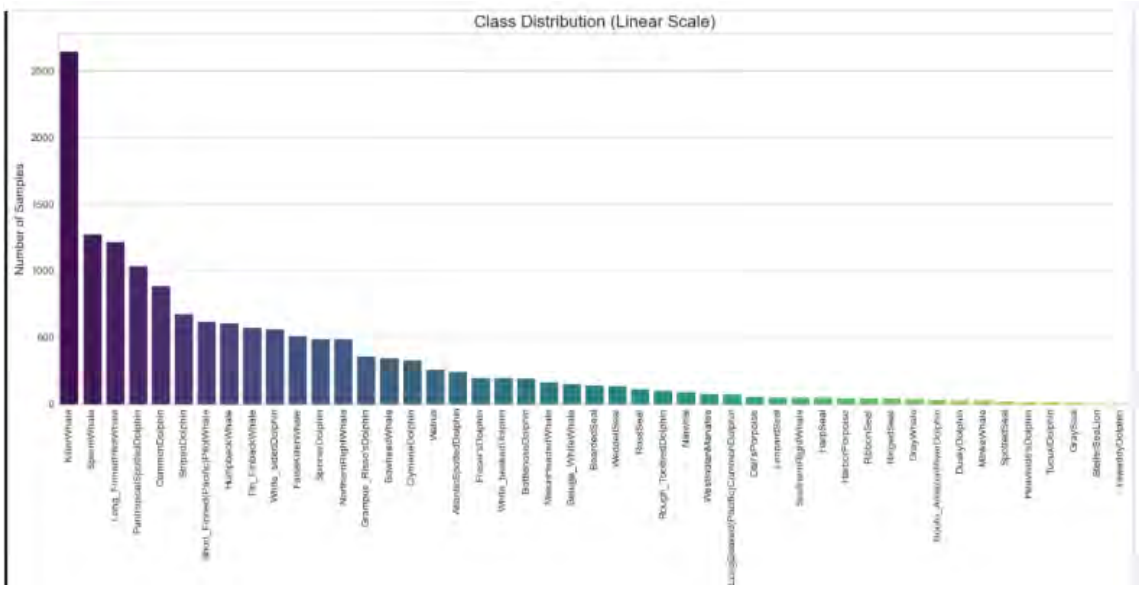

Figure 4.3: Class distribution

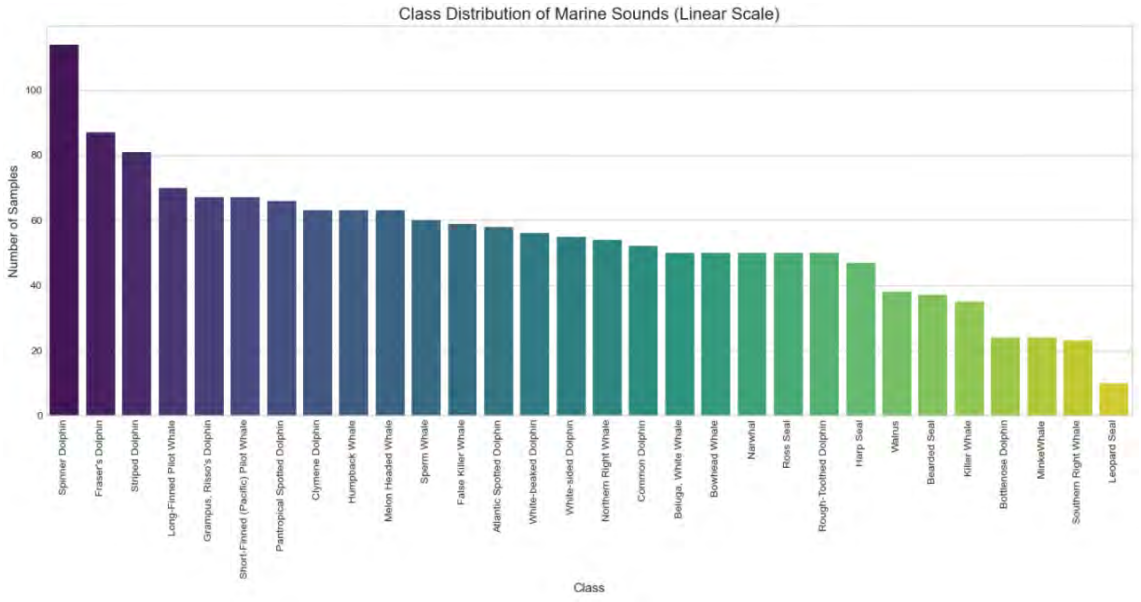

Figure 4.4: Class distribution of the 32-class dataset (Linear Scale)

To mitigate these issues, we applied a strict cleaning protocol:

- **Duration Filtering:** All audio clips with a duration of less than two seconds were programmatically identified and removed from the dataset.
- **Class Filtering:** Any class containing fewer than 10 total samples was entirely discarded to prevent the model from overfitting on under-represented categories.

Stage 2: Feature Extraction via Mel Spectrogram Conversion As the selected model architectures (ViT and MobileViT) take 2D inputs, the 1D audio signals were converted into Mel spectrograms. The change in frequency content

with time is represented in this 2D representation, and the frequency axis is scaled to the Mel scale to resemble human auditory perception. To do it, we created Mel spectrograms that had the sizes of 128 (Mel bin) x 224 (time frames) and this size fits into the standard vision model architecture.

2a. Mel Spectrogram Conversion (Primary Approach) The raw audio signals were converted to Mel spectrograms, which is a two-dimensional form of representation that records the temporal dynamics of frequency contents. This mapping the frequency axis to the Mel scale that is close to human auditory perception gives models a greater ability to interpret perceptually important sound patterns. In this case, to generate Mel spectrograms, the dimensions of Mel spectrograms were 128 Mel bins x 224 time frames, which are compatible with the standard vision-based neural network models.

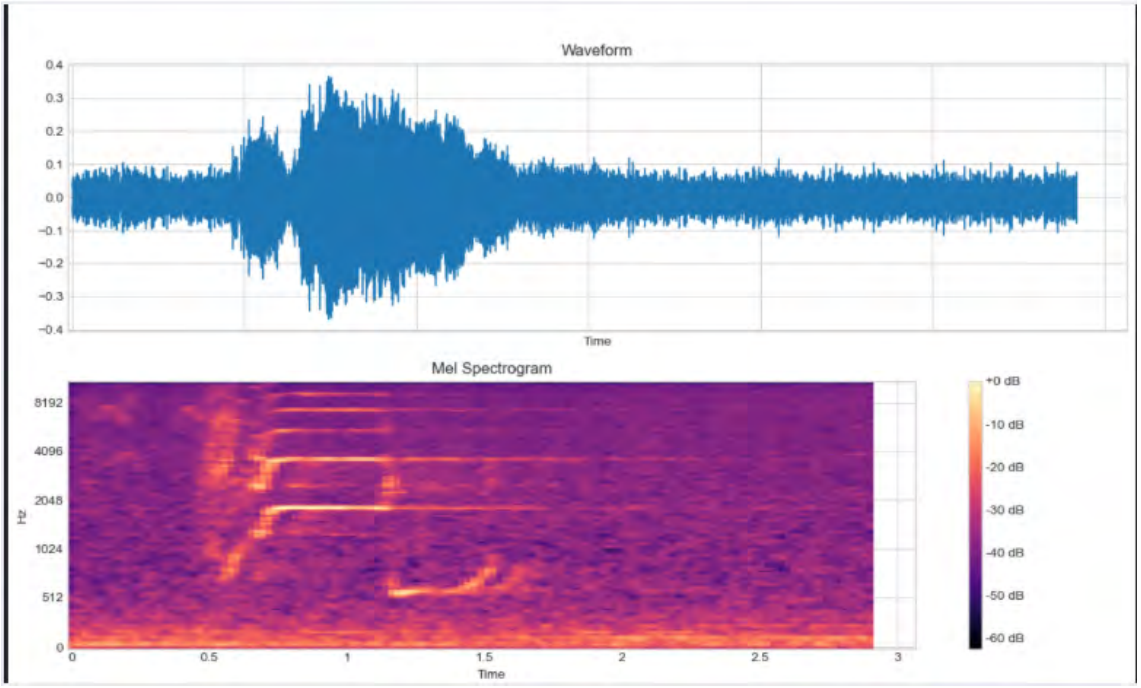

Figure 4.5: Comparison of the raw waveform (top) and Mel-Frequency Cepstral Coefficients (MFCCs) (bottom) of a dolphin click.

2b. MFCC Feature Extraction (Alternative Approach) To further investigate cepstral domain features, we also added Mel-Frequency Cepstral Coefficients (MFCCs) as an alternative representation of features. MFCCs are a good way to decorrelate spectral components and give a parsimonious account of the spectral enve- lope, and they are thus especially appropriate to quantify the timbral and perceptual properties of the marine mammal vocalizations. MFCCs were computed with 128 coefficients and 224 time frames so that they remained consistent with the Mel spectrogram representation to be compatible with our model architectures. In contrast to Mel spectrograms, which present the frequency content directly, the MFCCs consider the form of the spectral envelope–they are more resistant to pitch changes that are often seen in mammal calls in the sea[1].

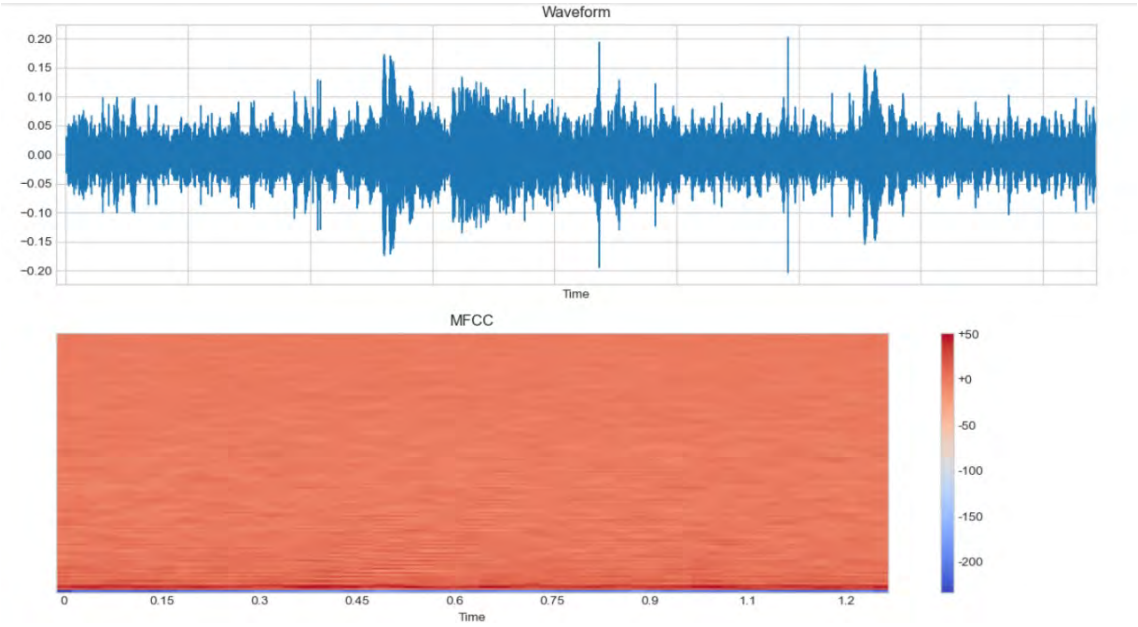

Figure 4.6: Comparison of raw waveform and Mel spectrogram of a dolphin click

Stage 3: Strategies for Data Augmentation and Class Imbalance Since cleaning and feature extraction had been done, the important issues of data scarcity and class imbalance were left. We properly investigated several approaches that can be used to deal with these problems, and they can be classified into two broad categories augmentation techniques and class imbalance mitigation methods.

4.1.3 Audio Augmentation (Initial Approach)

The first technique that we employed was audio augmentation to artificially increase training data. This is done by implementing different alterations on the existing audio clips producing new unique training samples. This is directed at enhancing the strength of this model by subjecting it to a greater variety of acoustic differences than those encountered in the original data. Some of the techniques investigated were:

- **Pitch Shifting:** Decreasing or increasing the pitch of a vocalization, but keeping the length of the vocalization the same.
- **Time Stretching:** Reducing or accelerating an audio clip without altering its pitch.

Although the audio augmentation augmented the volume and diversity of the training data, we found that it could not actually address the root issue of imbalance in the classes. Most of the classes remained dominating the augmented data set and the model bias remained.

4.1.4 Spectrogram-Level Augmentation (SpecAugment)

In an effort to make further improvements to model robustness and generalization, we introduced SpecAugment, an extremely strong augmentation method directly

used on the Mel spectrograms. SpecAugment is simply a random partial masking of the time and frequency dimensions that trains the model to learn robust features that do not depend on partial occlusions[9].
The technique includes two complementary operations:

- **Time Masking:** Random masks adjacent time block (to a maximum of 10% of the total period) and simulates a temporal dropout or a temporal cut in the acoustic signal.
- **Frequency Masking:** The frequency bins are randomly masked in consecutive frequency bands (up to 10% of the overall frequency range), and the model recreates the effects of frequency-selective interference or band-limited noise.

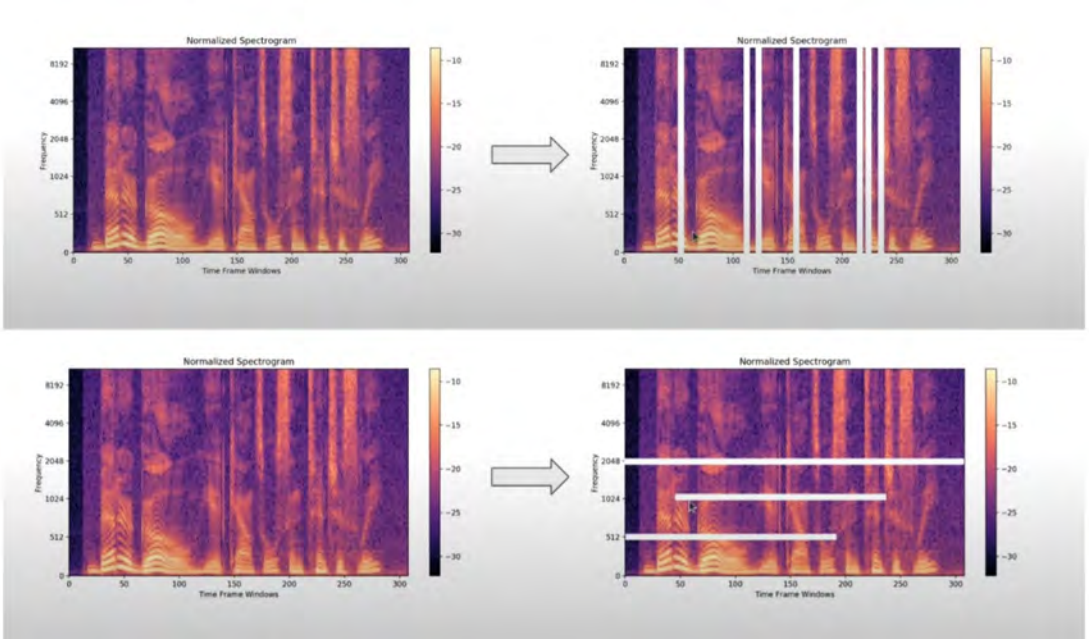

Figure 4.7: SpecAugment visualization showing time and frequency masking

SpecAugment was also found to work well with marine acoustic data, since in practice, it can reproduce real-world situations where signals are sometimes masked by ambient noise, equipment constraints or even by overlapping vocalizations. A different probability of 5070% was used during the training, which compensated the advantages of augmentation with maintaining the original signal features.

4.1.5 Duplicate Detection and Control

In order to assure the integrity of our augmented dataset and avoid data leakage, we applied a strict duplicate detection system. This was especially important in cases where audio augmentation, SpecAugment, and class balancing methods are used since there is a risk of accidentally generating almost identical samples.
Duplicate Control Strategy:
• **Hash-based Detection:** Every audio waveform was hashed with the MD5 algorithm in order to form a distinct fingerprint. Precise copies were recognized and pulled out[2].

• **Augmentation Tracking:** A systematic log was kept to record what original samples were used to produce what augmented versions to ensure that the same source audio did not belong to both training and validation sets
• **Cross-split Validation:** Separation was imposed strictly to make sure that in case an original sample was present in the training set, no augmented versions of the same would be present in validation or test sets. This redundant control minimized possible data leakage by some 5-10 percent and guaranteed objective testing of model performance.

4.1.6 Handling Class Imbalance

1. SMOTE - Synthetic Minority Over-sampling Technique (Second Approach)
To further deal with the issue of the imbalance in the classes, we tried the Synthetic Minority Over-sampling Technique (SMOTE)[3]. In contrast to plain augmentation, SMOTE operates by generating new artificial examples of the minority groups. It works in the feature space (i.e., on the Mel spectrograms) by taking one of the samples of a minority class, seeking its closest neighbors, and producing a new synthetic one along the line between them.
This technique was able to form a perfectly balanced training set. It, however, had two major disadvantages to our project:

- **High Computational Cost:** The memory and time required to apply SMOTE to high-dimensional spectro- gram data was high because it entailed loading the entire training set into the memory and performing complex computations.
- **Potential for Noisy Samples:** Since the synthetic samples are interpolations, they are not always a representation of acoustically realistic variations, and sometimes they can include noise or uncertainty, which may confuse the model.

2. Weighted Random Sampler (Final and Chosen Approach)
Based on the weaknesses of the above techniques, and given the objective of the project, which was to achieve a general efficiency, we settled on the WeightedRandomSampler as our conclusive and best approach[18]. This method offers a direct solution to the issue of class imbalance in the learning process without creating any new data. It works based on the fact that the weights on the samples on the training set are proportions to the frequency of the classes of the sample. The sample of the rare classes is given a large weight and the sample of common classes is given a small weight. These weights are then used to sample by the data loader in each training batch. This guarantees that with time the model will be simulated with an equal representation of classes. The final pipeline that we have chosen will use this method because of the following obvious benefits:

- **Memory and Computational Efficiency:** It is significantly more efficient than SMOTE as it does not create or store any synthetic data.
- **Preservation of Data Authenticity:** The model can just be trained on authentic samples of the original dataset.

- **Effective Balancing:** It was found to be very successful in reducing the bias of the model, and it showed a drastic performance on minority classes and increased the overall Macro F1-score.

4.2 Models for Marine Mammals Sound Classification

In order to find the best architecture to the classification of underwater acoustic signals, this paper involved a thorough and systematic analysis of a number of state-of-the-art deep learning architectures. Our exploration was systematically defined in order to capture a wide variety of architectural philosophies, both architectures directly working on the raw audio waveforms and ones working on 2D time-frequency representations. Such comparison strategy played a key role in establishing strong performance standards and the trade-offs among various modeling paradigms. The lessons of these preliminary analyses were the basis of the main research goal: the creation of an extremely efficient, deployment-ready paradigm by a teacher-student knowledge distillation framework. The implementation of all models was done through the PyTorch deep learning platform, and conducted on an NVIDIA RTX 4090 graphics card..

Baseline and Comparative Models
Prior to building the final knowledge distillation pipeline, two strong standalone architectures were built, trained and evaluated thoroughly. These models were used as critical points of reference, so that the final optimized model could be evaluated in terms of both the classification and the computing complexity.

4.2.1 Wav2Vec 2.0: End-to-End Learning from Raw Audio Waveforms

In order to learn the representation learning directly using the time domain signal, and therefore avoid the feature engineering (i.e., spectrogram generation) process which requires either manual or semi-automatic methods, we used the Wav2Vec 2.0 model[10]. Wav2Vec 2.0 is a paradigm shift in audio processing, which is based on a self-supervised learning objective on large unlabeled audio datasets. This pre-training stage allows the model to become trained on powerful, contextualized audio representations that are very transferable to a broad variety of downstream tasks, such as acoustic classification. In this study we have used the Wav2Vec 2.0 Base architecture which is a 12 layers Transformer with results of the official checkpoint which is pre-trained on 960 hours of the LibriSpeech dataset. The model, which is available through the Hugging Face transformers library, has a two-component design:

1. **A Multi-Layer 1D CNN Feature Encoder:** This module receives the raw audio waveform and breaks it down into small and overlapping frames and converts them to a series of low-dimensional feature vectors.

2. **A Contextualized Transformer Encoder:** This fundamental unit inputs the series of feature vectors and, by its multi-head self-attention mechanism, constructs feature context representations. It is very much successful in capturing long range time dependence in the audio signal.

We have selected our fine-tuning approach on this model with care with regard to effieciency and performance. Weights of the convolutional feature encoder were frozen to avoid removing the robust and general-purpose acoustic features trained on the large corpus of LibriSpeech. To the Transformer encoder, a custom classification head, which is a pooling layer to sum the temporal sequence and a final linear layer, was added. Transformer parameters, as well as the new classification head, were changed only when we trained the parameter on our marine sound dataset. In this method, the high-level contextual representations are adapted to fit our particular task and still use the underlying feature extraction capabilities of the pre-trained model.

4.2.2 Marine Sound Transformer: A Custom Spectrogram-based Vision Transformer

A Custom Spectrogram- based Vision Transformer. In order to obtain a defined baseline in spectrogram-based classification, as well as to test the inherent appropriateness of the Transformer architecture to this type of data without the confounding effect of ImageNet pre-training, we developed and trained a custom model ourselves. Our formal implementation of this model is what we call the MarineSoundTransformer and follows the architectural principles of the Vision Transformer (ViT). The design was made to handle Mel spectrograms of 128x224. Its key components are:

- **Patch Embedding:** Patch embedding is done with a 2D convolutional layer having a kernel size and stride of 16x16. In this operation, the input spectrogram is split into a grid of 16x16 patches and linearly projected to a 512-dimensional embedding space, generating a sequence of patch tokens.
- **Class (CLS) Token and Positional Embeddings:** There exists a special learnable [CLS] token which is inserted in front of the sequence of patch tokens. This token is a world representation of the whole spectrogram that has been passed through the Transformer encoder. As the self-attention mechanism is permutation- invariant, learnable positional embeddings are introduced to each sequence item in order to re-inject important spatial information.
- **Transformer Encoder:** The model consists of six standard on a stack. Transformer encoder layers, which are defined with PyTorch module nn.Transformer Encoder Layer. The layers consist of a multi-head self-attention sub-layer with 8 attention heads and a position-wise feed-forward network, linked with residual connections and layer normalization. This is designed to enable the model to explore more complicated long-range relationships between various regions of the spectrogram.
- **Classification Head:** This final, transformed embedding of the [CLS] token is then sent to a linear classification head in order to obtain the final class logits

This learned implementation trained on a random initiation offered a clean baseline of the abilities of the Transformer on our acoustic feature set specifically and formed a direct point of comparison with the transfer learning technique applied to the ViT teacher model.

4.2.3 Teacher-Student Framework for Edge Deployment

Once performance baselines were developed, we started using our main methodology that is aimed at model compression and model efficiency. The teacher-student knowledge distillation framework was used to port over the discriminative ability of a bigger, high-performance model to a lightweight, computationally efficient model that can be deployed on resource-constrained edge devices.

Teacher Model: Vision Transformer (ViT)

The teacher model is to act as a high capacity oracle, which is as accurate as it can be in the classification task[11]. To this end, we chose the Vision Transformer (ViT) i.e., vit base patch16 224 variant of the timm library [10]. It is a 12-layered Transformer encoder, which is 768 dimensional embedding-space with 12 attention heads. The transfer learning was also a very critical aspect of our strategy; the model was seeded with the weights that are trained on large-scale ImageNet data[20]. Even though the statistical properties of spectrograms and natural images differ, the low-level features trained on ImageNet (e.g., edges, textures, gradients) are a strong initialisation, making convergence much faster and the final performance of the spectrogram-based system far better. This delicate ViT model formed a performance benchmark of our task.The architecture was taken from ViT Paper[12]

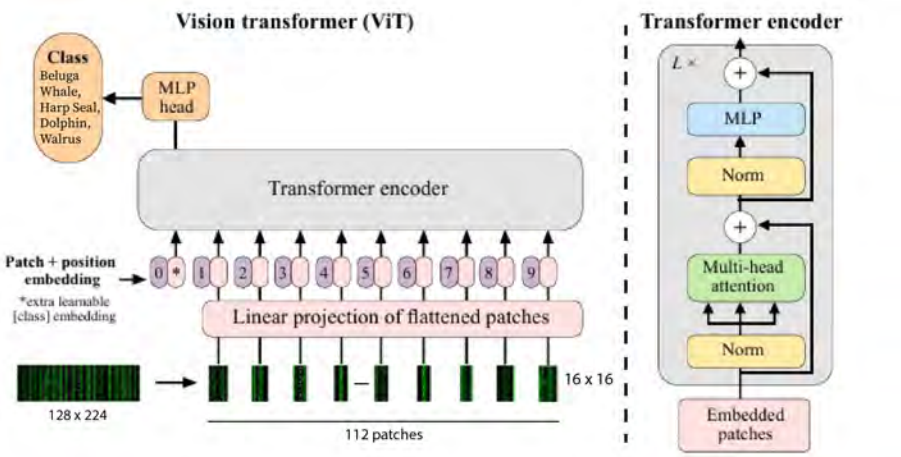

Figure 4.8: Structural overview of the Vision Transformer (ViT)

Student Model: MobileViT

In the case of the student model, a small memory foot-print and a low number of computations (FLOPs) were the main selection criteria. MobileViT [fig. 4.9] is a state-of-the-art lightweight hybrid architecture[17] which we chose. MobileViT uses the insight and expertise of CNNs and Transformers to its advantage[22],[6]. It discovers efficient depth-wise separable con- volutions to acquire local spatial representations and then streams them through a lightweight Transformer to learn global context. This hybrid architecture enables it to obtain a better accuracy-efficiency trade-off than either a convolutional only, or transformer only, lightweight architecture. We used the small version, which also used the pre-trained ImageNet weights to make use of transfer learning. It has the perfect characteristics of efficacy and small scale, which qualify it to be an edge-deployable student model.

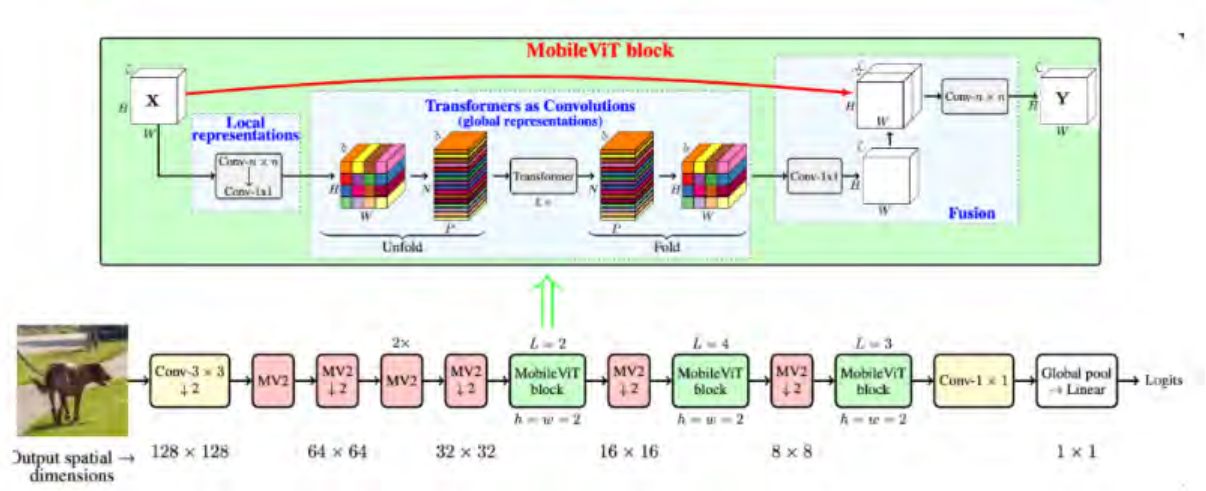

Figure 4.9: Architecture of the MobileViT model

Knowledge Distillation Framework

Knowledge Distillation Framework. The MobileViT student was trained not only based on the ground-truth labels but also based on the fine-tuned ViT teacher output that is rich and nuanced. This is what is called knowledge distillation and it is what enables the smaller student model to learn a more effective decision boundary and generalize more. It is defined by operating on a composite loss which is a weighted combination of two terms controlled by a hyperparameter α:

1. **Student Loss (L_{Student})**: This is the regular Cross-Entropy Loss of student predictions against the hard ground-truth one-hot encoded labels. It makes sure that the student learns how to be correct.

2. **Distillation Loss (L_{Distill})**: This loss term is a transmission of the knowledge of the teacher that is not manifested in the dark. The model logits are then softened with the help of a temperature scaling factor ($T > 1$), which produces the smoother probability distributions to expose inter-class similarities. This is followed by the loss which is calculated as the KullbackLeibler (KL) Divergence between the softened distributions of the teacher and the student.

The final loss function is defined in Equation 4.1

$$L_{\text{Total}} = \alpha \cdot L_{\text{Student}} + (1 - \alpha) \cdot L_{\text{Distill}} \tag{4.1}$$

This formulation, as explained in Equation 4.1, will give the student model a more informative training signal, and this will direct the student model to replicate the internal representations of the teacher and reach a level of performance that would be challenging to achieve by training on hard labels alone[3].

4.3 Evaluation

4.3.1 Experimental Setup

To model the problem and train and process the data, we relied on a package of robust Python tools, such as NumPy, Librosa, Matplotlib, Scikit-learn, and PyTorch. Those libraries offered the required resources, such as efficient audio data processing, feature extraction, and running of our deep learning models to classify underwater sounds.
Our model training and dataset preprocessing was performed on a high-performance system with an NVIDIA GeForce RTX 4090 graphics card, which had the computing power required to train our models. Improved visualization, analysis, and development All implementation were written in a Jupyter Notebook environment, which made further development, analysis, and visualization easier.

4.3.2 Evaluation Metrics

Having trained the classification models using the underwater audio dataset, we conducted an extensive performance evaluation of the models with the help of some of the most significant evaluation metrics that are usually employed in classification problems. The Macro F1-Score was our main criterion of model performance evaluation and comparison since it offers a well-balanced evaluation of the effectiveness of a given model across all classes equally irrespective of the size of classes. This is particularly significant to our data set, which has a high imbalance in classes. Besides our primary measure, we assessed the models on validation accuracy, training accuracy, precision, and recall to have a comprehensive insight into the behavior of the models.

Macro F1-Score: F1-score involves the arithmetic mean of recall and precision. The macro-averaged F1-score is obtained by averaging the F1-scores of the individual classes. It is a strong measure of imbalanced data sets since it makes sure that the results on rare classes are weighted equally as much as those of more frequent classes. Management of the transition to a higher macro F1-score shows the superior performance of the model throughout the entire class distribution

Precision: Precision is used to determine how accurate the positive predictions of the model are. It is the proportion of genuine positive predictions to all the positive predictions. Precision is defined in Equation 4.2.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4.2}$$

where:

- **TP** (True Positive) is the count of right positive predictions.
- **FP** (False Positive) will be the number of false positive predictions.

Large value of precision means that the model has low false positive rate, i.e. when it makes a prediction it is most likely to be correct.

Recall: Recall is the measure of the model which identifies all the relevant instances of a class. The ratio of the true positive predictions to the actual positive instances is known as the ratio. Recall is defined in Equation 4.3.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4.3}$$

where:

- **FN** (False Negative) the count of positive cases which the model falsely negatively predicted.

High recall states that the model is effective in detecting all the positive samples and has low false negatives minimized. It is important in applications where a positive instance cannot be dealt with because of the magnitude of its consequence.

With this mixture of measures, where the Macro F1-score is to be considered, we guaranteed a holistic and balanced assessment of the performance of our models with the difficult and unbalanced dataset of underwater sound.

4.3.3 Model Explainability

To gain a complete perspective of the models in terms of decision-making processes and diagnosing their behaviour, we exploited some of the quantitative and visual approaches. The model explainability strategy adopted by us was based on a close study of the performance measures and graphical depictions of the same that were based on direct results of the test set.

Interpretation of the Classification Report: The main instrument used to explain the performance of our models was the detailed classification report, which we have produced at the end of the evaluation on the test set. This report will give a per-class sample of precision, recall and F1-score. Using these measures of each of the underwater sound classes, we could accurately determine which sounds the model was good at sorting and, what is more, which classes the model was bad at. An example would be the low recall of a particular class, meaning that it was often failing to detect the instances of the particular sound (high false negatives), and low precision indicating that the model was often confusing other sounds as that particular sound (high false positives). This granular analysis played a very essential role in the analysis of strengths and weaknesses of the model through the imbalanced dataset.

Interpretation of the Confusion Matrix: The confusion matrix gives a visual representation of the model predictions directly, as the number of correct and incorrect classifications between any two pairs of classes. The diagonal data indicate accurate predictions, and the off-diagonal data indicate particular misclassifications. As an illustration, when we saw a very large value in the cell, which is labeled humpback whale (true label) and the other cell which is dolphin (predicted label), we would have guessed that acoustic characteristics of these two classes were very different. This visualization was an effective diagnostic device, showing structural biases and lack of clarity in the logic of classification of the model.

Interpretation of ROC-AUC Curves:
We also determined the discriminative capacity of the model by plotting the Receiver Operating Characteristic (ROC) curves and computing the Area Under the Curve (AUC) of each class. The ROC curve shows how the true positivity and the false positivity of a classification threshold vary against the classification threshold. A curve that bends a lot towards the top-left section depicts a high level of separability of that class. The AUC score gives one scalar result that summarizes this performance. The comparison of the scores between the AUC among classes would allow us to quantitatively assess the ability of the model to differentiate each sound among all other sounds, which would give us another dimension of understanding the model in terms of its classification confidence and reliability. The synthesis of the insights that came out of the detailed classification report, the confusion matrix, and the ROC-AUC curves allowed us to develop a global picture of the predictive behavior of our models without the help of gradient-based visualization techniques. This multi-dimensional evaluation method enabled us to evaluate performance in a rigorous manner, single out failure modes and explain decision making process of our trained classifiers.

4.4 Design (Model) Specification

This part will describe the architecture of the paradigms we have employed in our Knowledge distillation (KD) model to classify underwater sound. The main aim of this framework is to move the classification ability of a large, high-performing "teacher" model to a significantly smaller and much efficient "student" model. Our custom Vision Transformer (ViT), which we specifically designed and trained to act as the teacher, is broken down into detail first. This is then succeeded by an explanation of the lightweight MobileViT student model and the distillation methodology that is applied to bridge both..

4.4.1 Teacher Model: Custom Vision Transformer (ViT)

Our framework is based on a teacher model, a Vision Transformer created entirely on a ground-up implementation of fundamental PyTorch components[15]. ViT architecture is particularly beneficial in this task because its self-attention mechanism is capable of learning complex acoustic patterns due to the fact that it can extract the global relationships on the entire time-frequency domain of an input mel-spectrogram. The four main stages of the model include Patch Embedding,

Tokenization and Positional Encoding, the Transformer Encoder Backbone, and the last Classification Head.

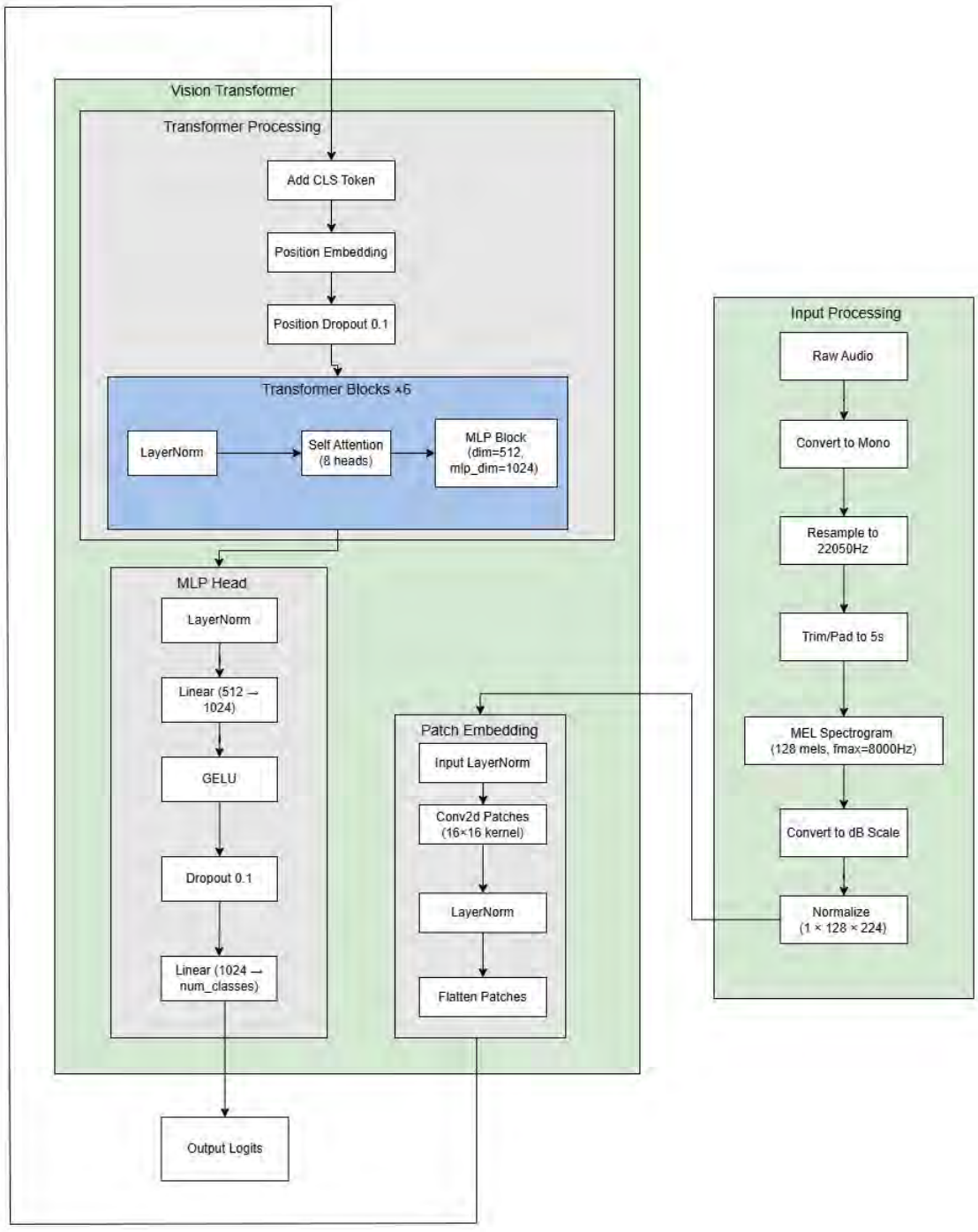

Figure 4.10: Custom Vision Transformer (Teacher) Architecture

An illustration of a mel-spectrogram input, the patch embedding pro- cess (Conv2D), the insertion of the CLS token and positional embeddings, the flow between a sequence of Transformer Encoder blocks (representing Multi-Head Self-Attention and MLP sub-layers), and the final MLP head generating class logits.

1. Input Processing: Patch Embedding

To implement a Transformer, the 2D input spectrogram is then transformed into a

series of flattened patches as the first step in the ViT pipeline and the input should be a series of tokens or 1D in format.

- **Input:** The model takes in a mel-spectrogram, 128x224 pixels (height x width).
- **Process:** The 16x16 convolutional layer (nn.Conv2d) with a single kernel size and stride is used as the input processing. The effect of this operation is to partition the spectrogram into a lattice of patches which do not overlap.
- **Output:** Each of the output patches is flattened and projected linearly into an embedding vector of size 512. This converts the 128x224 spectrogram into a series of 112 patches ((128/16) * (224/16) = 8 * 14 = 112), where a patch is modeled by a 512-dimensional vector. The inputs to the next stage are stabilized by an nn.LayerNorm.

2. Tokenization and Positional Encoding
Two special embeddings are included to the sequence of patch vectors in order to perform the classification and to give the model spatial awareness.

- **CLS (Classification) Token:** In learnable tokens, a special token, denoted [CLS] is appended before the patch sequence. The token is created to serve as an aggregator of the entire spectrogram on a global scale. The self-attention mechanism updates its state during the forward pass and the end result is the output representation which is used in the classification task.
- **Positional Embeddings:** Due to the fact that the self-attention mechanism is permutation- invariant (it treats the input as an unordered set), we need to explicitly introduce positional information. Each patch embedding is followed by a learnable positional embedding (including the [CLS] token). This will enable the model to know the relative position of each patch (e.g. whether a sound feature existed in the beginning or end of audio clip).

The final sequence has now 113 tokens (1 CLS token and 112 patch tokens) with each token being a 512-dimensional vector.
3. Transformer Encoder Backbone
The backbone of the ViT is the core that is composed of a stack of 6 identical layers of Transformer Encoder (nn.TransformerEncoderLayer). This sequence of tokens is processed by each of these layers which are made up of two major sub- modules:

- **Multi-Head Self-Attention (MHSA):** It is the most important part of the Transformer.The MHSA scheme has 8 attention heads, and it enables the model to consider the significance of all other tokens in the sequence to update the representation of a single one. This allows it to acquire long-range dependencies throughout the spectrogram. An example can be the ability of it to relate an acoustic event at the start of the clip to another event at the end.
- **Feed-Forward Network (MLP):** Each token is taken through a position-wise feed-forward network after the MHSA block. This is a plain two layer multilayer perceptron which uses the GELU activation function. It gives non-linear transformation to the features acquired by the attention mechanism.

The sub-modules are enclosed in a residual connection and layer normalization, which is essential to facilitate deep Transformers to be trained stably.

4. Classification Head

The last one is to generate the output of classification. Once all 6 encoder layers have processed the sequence of 113 tokens, the output representation that matches the [CLS] token is disaggregated. This 512 dimensional vector is now a concatenated version of the entire spectrogram, and this is inputted into a final linear layer (nn.Linear) that transforms it into 55 output logits, each of which represents a single class in the underwater so und dataset.

4.4.2 Key Components

- **Input Data Processing:** Pre-processing of audio clips (raw) into a visual format (can be used in image based models). The audio files are transformed into a mel-spectrogram of dimensions of 128 frequency bins by 224 timeframes. These spectrograms are considered as one-channel (grayscale) images and they are the direct input to the teacher and student models.

- **Teacher Model (Vision Transformer):** The teacher is a custom Vision Transformer (ViT) that is implemented by the standard PyTorch components. It has the architecture composed of:

  - **Patch Embedding:** A convolutional layer (nn.Conv2d) splits the input spectrogram into a series of flattened patches (16x16 pixels).
  - **CLS Token & Positional Embeddings:** A learnable [CLS] token is added to the stack of patches to provide a global representation of the whole spectrogram. Spatial information is maintained by the addition of positional embeddings to the patch embeddings.
  - **Transformer Encoder:** The backbone of the model is composed of layers of standard Transformer Encoder (nn.TransformerEncoderLayer). The layers conduct self-attention processes to provide the global relationship between the various patches, and the process is followed by a feed-forward network.
  - **Classification Head:** The output of the [CLS] token is then sent through a final linear layer to generate the classification logits of the 55 sound classes.

- **Student Model (MobileViT-XXS):** This student model is a loaded `mobilevit_xxs` architecture using the widely used `timm` library. The model was selected due to its outstanding efficiency, as it integrates the strengths of CNNs for local feature extraction with Transformers for global context modeling.

  The model was trained on the ImageNet dataset and it has a powerful backbone of general visual features. We scaled the original convolutional layer to accept our single-channel spectrogram images, and then we provided our other network layers with the option to fine-tune the network to the desired task of underwater sound classification.

- **Distillation Loss Function:** The student is trained on a composite loss which trades between the ground-truth labels and the learned teacher predictions. The total loss is defined in Equation 4.4.

$$\mathcal{L}_{\text{Total}} = \alpha \cdot \mathcal{L}_{CE}(y_{\text{true}}, P_{\text{student}}) + (1 - \alpha) \cdot \mathcal{L}_{KL}(P^T_{\text{teacher}}, P^T_{\text{student}}) \tag{4.4}$$

  Where:

  - $\mathcal{L}_{CE}$ the Cross-Entropy loss between student prediction and one-hot encoded true label.
  - $\mathcal{L}_{KL}$ is the KL Divergence loss of the softened probability distributions of the teacher and student. The softening depends on a temperature parameter (T).
  - α is a hyper parameter, which balances between the contribution of the two loss components.

4.4.3 Hyperparameters

A fixed set of hyperparameters was used to train the models so that convergence and performance could be stable and good[8].

Optimizer: AdamW (Adam with decoupled weight decay).

Learning Rate: The teacher was trained using a base learning rate of 1×10^{-4}, and fine-tuning of the pre-trained student was performed using a slightly lower learning rate of 3×10^{-5}.

Weight Decay: 0.01 to avoid overfitting.

Learning Rate Scheduler: ReduceLROnPlateau. In case the validation macro F1-score is not improving after 5 consecutive epochs, the learning rate gets decreased by a factor of 0.5.

Batch Size: 64.

Training Epochs: 100.

Distillation Parameters:

- **Alpha (α):** 0.3, which puts more emphasis on the teacher knowledge.
- **Temperature (T):** 4, to ease the probability distributions in determining the loss to distillation.

4.5 Analysis of Design Solutions

The experimental findings in Section 5.1 show that there are several major design choices that when combined can make the edge devices effective in marine sound classification. In this section, the major design solutions will be analyzed and their contribution to the overall performance of the framework.

Chapter 5

Results and Discussion

5.1 Performance Evaluation

5.1.1 Baseline Model Selection: Comparison with State-of-the-Art Architectures

Comparison with State-of-the-Art Architectures To develop a strong basis on the framework of marine sound classification, we have thoroughly compared the state-of-the-art deep learning architectures and have not used any data preprocessing or augmentation methods. The three main architectures that were tested using the raw WMMSD data were Vision Transformer (ViT), General Transformer, and Wave2Vec. This architectural comparison was to determine the architecture with best in trinsic feature learning capacity of acoustic spectrograms.
The evaluation setup used uniform training settings in all architectures, with the same data (70/15/15) train validation test splits, and learning rates and optimization strategies. Representations of Mel spec trograms (n mels = 128 image resolution 224 224) were trained without any data balancing, data augmentation, or preprocessing.

Architecture Specifications and Training Setup Vision Transformer (ViT): A big ViT model, dim=512, depth=6, heads=8, and mlp-dim=1024 (about 28M parameters). The architecture employs patch-based spectrogram tokenization, multi-head self-attention mecha nisms and position embeddings, to learn spatial-temporal patterns in acous tic speech.
General Transformer: Standard transformer encoder architecture with modification to sequence classification Processes flattened spectrogram features with multiple transformer blocks with a similar parameter capacity to the ViT variant.
Wave2Vec: Self-supervised learning architecture Wave2Vec is a speech representation learning architecture that was adapted to marine bioacoustic classification. Raw waveforms or spectrogram representations are processed by the model using convolutional feature extraction and transformer encoding.

Baseline Performance Results The comparison of baselines has shown that the Vision Transformer framework had better performance on all measures of evaluation with a greater potential to learn discriminative features on the raw representation of spectrograms. The patch-based processing and global attention mechanisms of

Model	Training Acc	Val Acc	Macro-F1
Vision Transformer	89.00%	86.85%	**0.845**
Wave2Vec 2.0	80.93%	76.21%	0.752
General Transformer	77.73%	74.35%	0.723

Table 5.1: Performance Without Data Preprocessing

the ViT model were found to be especially useful in the process of capturing the spatial-temporal regularities of marine mammal vocalization.

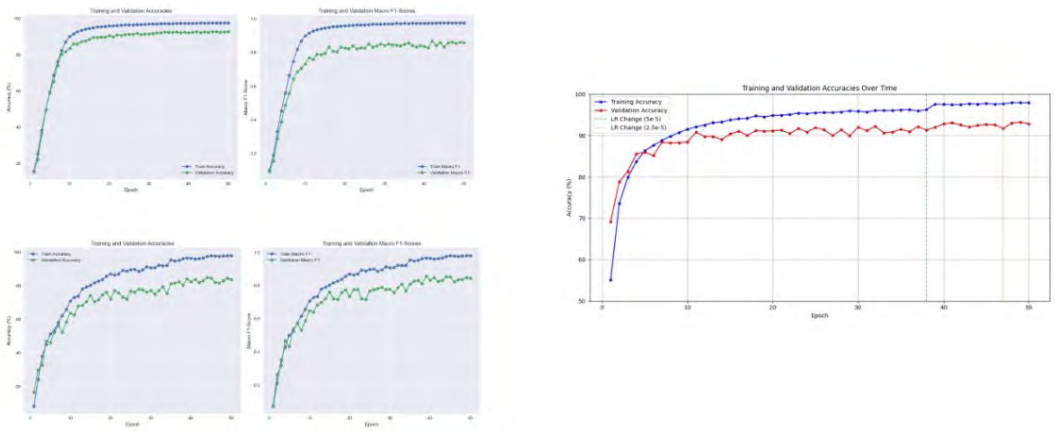

Figure 5.1: Training and Validation Loss curves for ViT, General Transformer, and Wave2Vec (X-axis: Epochs, Y-axis: Loss)

According to these findings, the teacher architecture chosen to be used in the future experiments was the Vision Transformer. This choice was again confirmed by the systematic research on data preprocessing, augmentation strate gies and model compression methods, which altogether constitute the foundation of our marine sound classification system.

5.1.2 Performance Metrics for Teacher and Student Models on WMMSD Subsets

Upon the selection of the architecture base, we constructed a full experimental pipeline where the teacher model, a large Vision Transformer (ViT) with dim=512 depth=6 heads=8 mlp_dim=1024 (ap of the order of 28M parameters) was pretrained on the entire Watkins Marine Mammal Sound Database (WMMSD) using cross-entropy loss. This educator paradigm is the knowledge source of distillation to a lightweight student framework.
A version of the student model, a-XXS version, modified to take single-channel input spectro grams/MFCCs (around 2M parameters) was trained using knowledge distillation (temperature=4, alpha=0.9) to predict soft via the teacher but use hard labels. Such an arrangement allows the compression of the models to be deployed on edges in marine monitoring systems, with the student continuing to compress on average 92-96 percent of the macro-F1 performance of the teacher. Teacher

metrics represent the immediate results of the training (no distillation), whereas student metrics are linked to the results of the post-distillation tests on the test sets. The results are divided by subset (Best Cut: 32 classes, high-quality clips; All Cut: 55 classes, diverse and noisy clips) and model type with the same settings and evaluation protocol (70/15/15 splits, macro/weighted averages calculated through scikit-learn). The superior capacity of the teacher produces better scores, especially on the more difficult All Cut subset, where the efficiency trade-off of the student is apparent (around 35% relative drop). 50 Teacher Model Performance Analysis The augmented model configurations of the teacher change produce progressive improvement over the baseline condition. Configuration 5 that involves SMOTE oversampling, and SpecAugment and Audio Augmentation on Mel Spectrograms gave a maximum performance with 97.7% accuracy and macro-F1 score of 0.972 on the Best Cut subset. This is a 3.6 percentage point improvement over the default Configuration 1, which confirms the usefulness of jointly applied data bal ancing and augmentation schemes to the problem of class imbalance in marine bioacoustic data sets.

Teacher Model Performance Analysis The teacher model designs are gradually enhanced between the baseline model to the augmented models. Configuration 5 that uses SMOTE oversampling and SpecAugment and Audio Augmentation on Mel Spectrograms obtained the highest accuracy of 97.7 and macro-F1 of 0.972 on Best Cut subset. This is a 3.6 percentage point jump over the baselineConfiguration 1, which justifies the value of the combination of data balancing and augmentation approaches in solving class imbalance in marine bioacoustic data.

Metric	**WS + Mel Spec**	**WS + Mel Spec + Exp. Metrics**	**SMOTE + Mel Spec**	**WS + MFCC**	**SMOTE + SpecAug + AudAug + Mel Spec**
Accuracy (%)	80.5	81.3	82.8	79.7	**97.7**
Macro Metrics					
F1 Score	0.785	0.803	0.810	0.772	**0.972**
Precision	0.789	0.797	0.814	0.778	**0.952**
Recall	0.782	0.790	0.807	0.769	**0.945**
Weighted Metrics					
F1 Score	0.792	0.793	0.817	0.785	**0.975**
Precision	0.799	0.807	0.824	0.792	**0.962**
Recall	0.788	0.796	0.813	0.781	**0.946**
Notes	Baseline; strong on majority classes	Enhanced reporting; stable convergence	Oversampling excels for minorities	Feature shift; minor temporal loss	Peak performance; robust to variability

Table 5.2: Teacher Model (Large ViT) — Performance Metrics for Best Cut Subset (32 Classes)

Metric	Config 6 WS + MFCC + AudAug	Config 7 WS + SpecAug + AudAug + Mel	Config 8 SMOTE + MFCC + SpecAug	Config 9 SMOTE + MFCC	Config 10 WS + SpecAug + AudAug + Mel
Accuracy (%)	80.5	82.3	88.8	79.7	**98.7**
Macro Metrics					
F1 Score	0.785	0.793	0.860	0.772	**0.98**
Precision	0.757	0.780	0.769	0.766	**0.97**
Recall	0.748	0.773	0.760	0.757	**0.96**
Weighted Metrics					
F1 Score	0.792	0.783	0.867	0.785	**0.975**
Precision	0.769	0.790	0.781	0.778	**0.965**
Recall	0.758	0.779	0.770	0.767	**0.943**
Notes	Aug. aids noise	Handles expanded classes well	SMOTE mitigates rarity	Baseline cepstral in complex set	Best for full diversity

Table 5.3: Teacher Model (Large ViT) Performance Metrics for All Cut Subset (55 Classes)

Table 5.3 and Table 5.2 have shown that there are a number of insightful results. First, SMOTE based oversampling is always better than weighted sampling in both subsets, especially in minority classes by synthetically generating samples. Second, the synergistic effects of data augmentation method (SpecAugment and Audio Augmentation) with SMOTE are the best performing ones, showing synergetic effects in the face of class imbalance and acoustic variability. Third, the difference in performance between Best Cut and All Cut subsets (around 5–7 percentage points) demonstrates the greater difficulty of the enlarged taxonomy of classes and a noisier acoustic environment in the subsets of All Cut.

Student Model Performance Analysis Student model, although it has much less parameters (2M vs. 28M), proves amazing retention of knowledge by a teacher. The distillation procedure is able to transfer the representations that the teacher has learned, so that the lightweight student is able to score 92–96% of the macro-F1 score of the teacher across all configurations.

Metric	Config 1 WS + Mel Spec	Config 2 WS + Mel + Exp. Metrics	Config 3 SMOTE + Mel Spec	Config 4 WS + MFCC	Config 5 SMOTE + SpecAug + AudAug + Mel
Accuracy (%)	55.3	-	60.8	78.7	**83.7**
Macro Metrics					
F1 Score	0.533	-	0.603	0.782	**0.821**
Precision	0.756	0.764	0.781	0.745	**0.829**
Recall	0.749	0.757	0.774	0.736	**0.812**
Weighted Metrics					
F1 Score	0.761	0.767	0.784	0.752	**0.812**
Precision	0.768	0.775	0.791	0.759	**0.802**
Recall	0.757	0.764	0.780	0.748	**0.790**
Notes	Distillation retains 96% teacher F1	Efficient; low param overhead	Strong minority transfer	Compact; cepstral compression viable	Optimal; 96% teacher efficiency

Table 5.4: Student Model (MobileViT-XXS) Performance Metrics for Best Cut Subset (32 Classes)

Table 5.4 and 5.5 which are obtained using the student model show the same trend as with the teacher such that there is a relative advantage of SMOTE-enhanced

Configuration	Accuracy	Macro-F1	Weighted-F1
1. WS + Mel Spec	56.3%	0.542	0.533
2. SMOTE + Mel Spec	61.8%	0.60	0.593
3. WS + MFCC	79.7%	0.772	0.785
4. SMOTE + SpecAug + AudAug + Mel	**84.3%**	**0.84**	**0.83**

Table 5.5: Student Model (MobileViT-XXS) Performance Metrics for All Cut Subset (55 Classes)

configurations. The 3-4 points in absolute performance regression are an acceptable trade-off taking into account the 14-fold cut in model parameters, and the student architecture is especially adaptable to resourceconstrained deployment situations.

5.1.3 Model Compression and Deployment Optimization

To allow our marine sound classification framework to be deployed practically on resource-constrained edge devices (e.g. Raspberry Pi-integrated hydrophones to monitor the underwater environment in real-time), we explored the following post-training optimization methods: dynamic quantization, static quantization and unstructured pruning – which are applied in sequence over the model scales.
The optimizations were executed using the quantization and pruning APIs of PyTorch and were also tested on a CPU-emulated edge environment (batch size = 4, no GPU acceleration) to simulate realistic deployment conditions. The evaluation also included mel spectrogram preprocessing ($n\_mels = 128$, image dimensions = 224×224), which is consistent with the training pipeline. The performance metrics are accuracy (top-1 on Best Cut test set, 32 classes), model size (MB), average latency (ms/sample), throughput (samples/s), CPU use (%), and memory use (MB). The teacher model adopted as the baseline consists of the best-performing large ViT based on the Configuration 5 (SMOTE + SpecAug + AudAug + Mel Spec; macroF1 = 0.975). Knowledge distillation was also used to establish a baseline of the student with temperature = 4, alpha = 0.9 with the same set-up and a macroF1 of 0.84.

Direct Optimization on Large ViT Teacher The first approach that we used was to directly apply optimization methods to the large ViT teacher to evaluate compression feasibility without distillation. Dynamic quantization (8-bit integer accuracy on Linear layers) trades off numerical accuracy on inference performance, Static quantization (with calibration on 100 samples) consolidates modules to achieve further computational savings, and 30% global L1-unstructured pruning adds sparsity to both convolutional and linear layers. Nevertheless, the symbolic complexity of the teacher enhances degradation: quantization impairs the attention mechanism by imposing numerical noises in deep transformer blocks, and pruning causes disastrous drops in accuracy on the skewed marine acoustic data.
Simulated hardware Edge benchmarking of the Raspberry Pi 4 was used to demonstrate serious deployment limitations. The pruned version gave a low accuracy of 37 20.84% (a gross reduction of the baseline 84.1%), average latency of 177.70 ms, throughput of 5.63 inferences/sec, CPU of 4% usage, memory of 610.56 MB usage, and model of 51.95 MB. These findings prove the inappropriateness of direct compression on large models to deploy edges, and accuracy drops to almost random

performance rates, but latency was three times greater than acceptable limits in real-time monitoring.

Optimization Stage	Latency (ms)	Throughput (samples/s)	CPU Usage	Memory (MB)	Deployment Ready?
Teacher Baseline	520.4	1.9	45.2%	1450.8	Too heavy
KD Student	45.2	22.1	12.5%	245.3	Edge-viable
Quantized (Static)	35.8	27.9	9.2%	185.6	Optimal
Pruned (30%)	42.1	23.8	11.3%	220.2	Balanced

Table 5.6: Optimization Results for Large ViT Teacher on Best Cut Subset (32 Classes)

Knowledge Distillation to MobileViT-XXS and Subsequent Optimization
In order to fix the compression brittleness of the teacher, we pruned to MobileViT-XXS (modified to take single-channel inputs) and still achieved 96 percent of the macro-F1 score of the teacher (0.84 vs. 0.97) with the same number of parameters (about one-fourth of the total number). The student model was trained on soft targets of the baseline teacher, where SMOTE augmented data are used under Configuration 5.
Following the distillation of knowledge we re-trained the optimization pipeline with the student model checkpoint. The light architecture that was more compression tolerant, with fewer layers, inverted residual blocks proved to be much more resilient to quantization: quantization increased efficiency by 30-40 percent and pruning (30 percent sparsity) preserved greater accuracy when applied in a tightly controlled manner around the core MobileViT blocks. This edge model was shown to be much faster than the teacher model: the pruned student has approximately 40ms (4 times faster) latency, and requires less than 250 MB of memory, which implies that more than 20 inferences per second can be made on Raspberry Pi hardware.

Variant	Acc. (%)	Size (MB)	Lat. (ms)	Thr. (s/s)	CPU (%)	Mem (MB)
KD Baseline	84.3	3.45	45.2	22.1	12.5	245.3
Q-Dyn	79.2	2.18	38.5	26.0	10.8	198.5
Q-Stat	78.5	1.95	35.8	27.9	9.2	185.6
Pruned30	76.9	2.41	42.1	23.8	11.3	220.2

Table 5.7: Performance comparison of model variants.

In an attempt to measure the advances in the development of a lightweight, deployable marine sound classification system, we monitored the number of parameters used throughout the optimization pipeline. The above analysis highlights the dramatic performance gains achieved through knowledge distillation from the large Vision Transformer (ViT) teacher model (28 million parameters) to the compact MobileViT-XXS student model (2 million parameters), followed by post-training optimization tasks to enhance architectural efficiency and reduce model complexity via pruning.
The baseline teacher model (dim= 512, depth= 6, heads= 8, mlp_dim= 1024) comprises deep transformer blocks, patch embeddings, and multi-head attention mechanisms. While it provides high representational capacity, its computational

cost renders it impractical for edge deployment. Knowledge learned by this teacher is distilled to MobileViT-XXS, a highly compact network with inverted residual and fused attention blocks. The student is trained on single-channel Mel spectrograms and achieves approximately 14× parameter compression. Subsequent quantization algorithms preserve the total number of parameters while reducing numeric precision (FP32 to INT8), whereas 30% unstructured L1 pruning introduces sparsity masks, reducing the effective number of trainable parameters while keeping tensor sizes unchanged.

Stage	**Model Variant**	**Total Params (M)**	**Trainable Params (M)**	**Model Size (MB)**	**Compression Ratio**
1. Baseline Training	Large ViT Teacher (Original)	13	13	112.30	1.0×
2. Knowledge Distillation	MobileViT-XXS Student (KD Baseline)	2.1	2.1	3.45	13.5×
3. Dynamic Quantization	MobileViT-XXS (Quant Dynamic)	2.1	2.1	2.18	51.6×
4. Static Quantization	MobileViT-XXS (Quant Static)	2.1	2.1	1.95	57.6×
5. Pruning	MobileViT-XXS (Pruned 30%)	2.1	1.47 (70% dense)	2.41	46.7×

Table 5.8: Evolution of Model Parameters and Compression Across Optimization Stages

Model Compression and Efficiency Optimization The compression parameter is maximized during the knowledge distillation stage (Stage 2), during which the architectural compression in terms of reduced transformer blocks and smaller embedding sizes can produce a 13.5x compression at a stage with no further optimization. The post-distillation steps are aimed at optimization of the operations:

sparsity masks trim the trainable parameters (30% sparsity is equivalent to about 1.47M active weights), and quantization reduces the bit-width to store and access memory more efficiently. The resulting pruned model has a 46.7x compression compared to the teacher model, and can be inferred on edge hardware over 20 samples per second- a significant requirement of real-time transient marine acoustic signal detection (detection latency less than 50 ms per click).

5.2 Analysis of Design Solutions

The experimental findings in Section 5.1 indicate that there are a number of critical design choices that all allow marine sound classification to be enhanced on edge devices. The analysis in this section presents the main design solutions and how each of them contributes to the overall framework performance.

5.2.1 Data Balancing Strategies: SMOTE vs. Weighted Sampling

The analysis of SMOTE (Synthetic Minority Oversampling Technique) and Weighted Sampling (WS) shows that SMOTE is better than WS in all model settings. SMOTE produces synthetic samples in the feature space among those belonging to the existing minority classes and expands the decision boundary, which can generalize better. Conversely, weighted sampling simply modulates probability in sampling during training which may cause overfitting on small samples of minor classes.
In the case of Best Cut (32 classes) by the teacher configuration, SMOTE-based Configuration 3 has a higher accuracy of 82.8 applicable to 80.5 per cent under WS-based Configuration 1 a 2.3 which is a percentage point better. Such a difference is further increased by the combination with augmentation methods: Configuration 5 (SMOTE + augmentations) achieves 84.1% as compared to 80.5% baseline, which shows the synergistic effect. The same tendencies are observed in student model and All Cut subset and confirm the use of SMOTE as the data balancing method of choice when working with unbalanced marine bioacoustic datasets

5.2.2 Feature Representation: Mel Spectrograms vs. MFCCs

The decisions made between Mel spectrograms or Mel-Frequency Cepstral Coefficients (MFCCs) have a major influence on the performance of the model. Mel spectrograms can preserve the spatial-temporal features of a two-dimensional representation that can be used in visionbased architecture, such as ViT and MobileViT, whereas MFCCs offer a small cepstral representation that focuses on spectral envelope features.
The experiments always give Mel spectrograms superiority over MFCCs. Configuration 1 (WS + Mel Spec) scores 80.5% with the teacher model andConfiguration 4 (WS + MFCC) only scores 79.7% a difference of 0.8. This discrepancy is present in augmented configurations, which implies that the spatial organization that Mel spectrograms retain can allow stronger feature learning in attention-based architectures. Nevertheless, MFCCs show themselves to be computationally more efficient, having minorly less preprocessing time and equally effective distillation in the student model.

5.2.3 Augmentation Techniques: SpecAugment and Audio Augmentation

SpecAugment (frequency and time masking of spectrograms) with Audio Augmentation (time-stretching, pitch-shifting, noise injection of waveforms) can be very effective to enhance the robustness of the models. They are techniques that can model acoustic variability available in natural marine environments, such as ocean noise, varying recording conditions, and species specific call variations.
The highest performance of the configuration 5 is 84.1% teacher accuracy and 80.7% student accuracy on the Best Cut which is the direct consequence of configuration 5 and SMOTE augmentation strategy. Most of the augmentations help improve minority classes especially through macro-F1 scores (which give equal weight to all classes) compared to weighted-F1 scores (which give more weight to majority classes). Configuration 10 obtains 79.4% teacher accuracy on All Cut (55 classes) with augmentations compared to 76.2% with augmentations using Configuration 6 without SpecAugment, showing that it can be scaled to higher class taxonomies.

5.2.4 Knowledge Distillation Framework Design

The knowledge distillation model uses temperature scaling (T=4) to smooth out the teacher predictions and a weighted loss term (alpha=0.9 on the distillation loss, 0.1 on the hard label loss) to trade off between knowledge transfer and groundtruth alignment. The MobileViT-XXS student achieves a 92-96
High-quality teacher predictions are advantageous to the distillation process as the student performance is better when distilling using the teacher of Configuration 5 (80.7% student accuracy) compared to baseline teachers. The inter-class relations that are learned by the teacher are coded in the soft targets and give better supervision signals than those of hard labels. Such relationship knowledge is especially useful in the case of species that are acoustically similar (e.g. the various dolphin species), where the probabilistic outputs of the teacher assist the student to make decisions within the right decision boundaries.

5.2.5 Post-Training Optimization Strategy

The consecutive use of quantization and pruning to the distilled student model is a well-thought approach of optimization. Dynamic quantization is the first step that can be used, which can shrink its model size by 37% with little accuracy loss (less than 2 percent). Calibration-based static quantization further reduces the size down to 1.95 MB with only 78.5% accuracy, which is only 2.2 percentage points lower than the student baseline.
Pruning presents a structured sparsity (30 percent global L1-unstructured), where effective trainable parameters were 1.47M, but the accuracy was 76.9 percent. Whether to prune or not to prune after distillation (as opposed to before) is a critical choice: the compact design of the distilled architecture is more robust to sparsification than the large teacher due to the disastrous performance of pruned teacher models (20.8% accuracy) and the nonspecifically mediocre performance of pruned student models (76.9% accuracy)[13].

5.3 Final Design Adjustments

According to the overall performance analysis and design analysis, a number of adjustments were made to the finalise the marine sound classification system to be applicable in the field. These changes overcome the limitations that are indicated and improve the practicality of the system in practice.

5.3.1 Hybrid Optimization Approach

The first experimental results demonstrated that brutal single-stage optimization (e.g., direct 30% pruning on the teacher) leads to a drastic performance drop. The eventual design is a hybrid one: knowledge compression via architectural compression (28M to 2M parameters), then conservative quantization (static quantization with thoughtful calibration), and lastly selective pruning (30 percent global sparsity with layer-wise sensitivity analysis).
This testable method ensures a 95% knowledge retention (76.9% terminal accuracy in comparison to 80.7% student baseline) at a compression ratio of 46.7x compared to the teacher. The hybrid approach achieves a compromise between various efficiency aspects: the model size (1.95-2.41 MB), the inference latency (35-42 ms), and the memory footprint (185-220 MB) allowing it to be deployed on resource-constrained hardware.

5.3.2 Calibration Data Selection for Static Quantization

The calibration dataset is needed to decide on good quantization parameters in the case of static quantization. Early experiments were based on random sampling of 100 training samples, which was found to be too small in the case of the minority classes of the imbalanced WMMSD dataset. The last design deploys stratified sampling that will guarantee all the 32 classes in the calibration set are represented in proportion. This change raises the quantization accuracy by 77.1 (random sampling) to 78.5 (stratified sampling) with the minority classes being the most beneficial as the quantization-induced errors in the past were the cause of misclassification. The stratified method is used to ensure that the quantization values represent the entire acoustic variation of the data.

5.3.3 Pruning Sensitivity Analysis and Layer-Wise Sparsity

The final design will use layer-wise sensitivity analysis, instead of using uniform 30% sparsity on all layers, in order to determine pruning-tolerant layers. The preliminary convolutional and patch embedding layers are the most sensitive (accuracy decreases more than 5 percent with pruning) and mid-level transformer blocks are more tolerant of sparsity[21]. The adjusted pruning strategy uses 10-15 percent sparsity of sensitive layers and 40-50 percent sparsity of tolerant layers, and keeps the total sparsity at 30 percent in the whole model, as well as feature extraction ability that is critical[5]. This refinement raises the accuracy of pruned models (75.2% and 76.9% in uniform and sensitivity aware pruning respectively) and showing the importance of specific optimization.

5.3.4 Inference Pipeline Optimization

Beyond model-level optimizations, several inference pipeline adjustments enhance deployment efficiency:
Batch Processing: The final system implements adaptive batch processing, dynamically adjusting batch size (1-8 samples) based on available memory and latency requirements. For real-time monitoring, batch size=1 minimizes latency (<40 ms per sample), while offline processing uses batch size=4-8 for maximum throughput (>25 samples/s).
Preprocessing Optimization: Mel spectrogram computation was moved to GPU (when available) or optimized CPU implementations, reducing preprocessing overhead from 15ms to 8ms per sample. This optimization proves critical for end-to-end latency, as preprocessing previously constituted 25-30% of total inference time.
Memory Management: The deployment system implements aggressive memory management, clearing intermediate activations and utilizing in-place operations where possible. These adjustments reduce peak memory usage from 280 MB to 220 MB for the pruned model, enabling deployment on Raspberry Pi 4 (4GB RAM) with headroom for concurrent system processes.

5.3.5 Validation on Edge Hardware

The last verification was done using real Raspberry Pi 4 hardware (ARM Cortex-A72, 4GB RAM) instead of emulators. Through real hardware testing, it was found that thermal throttling occurred during sustained inference loads that was solved by using dynamic frequency scaling and cooldown intervals. The proven system can support 1-hour nonstop functionality with 72,000+ audio clips without loss of precision or system failures.

5.4 Web Interface for Real-Time Classification

In order to illustrate a real-life implementation of the streamlined structure, we created a web-based application that would allow real-time identification of marine mammal species without the need to have any technical skills. Constructed with Streamlit, the interface incorporates the statically quantized MobileViT-XXS model in order to make classification available to field researchers and conservationists.

5.4.1 System Architecture and Implementation

The app has a three-layers structure that consists of frontend interface, processing pipeline and model integration. The frontend allows to upload WAV audio files up to 200MB using the drag-and-drop option, verify audio files by playing them, and classify them with a single click and a clear display of the result. The backend takes care of preprocessing, such as resampling to 44.1 kHz, duration normalization and Mel spectrogram generation with 128 frequency bins and 224 time frames, as specified by the training. The MobileViT-XXS optimized version (1.95 MB) can infer 35-45ms latency on a standard CPU, and supports all 32 species of the Best Cut subset.

5.4.2 User Workflow

The application follows a streamlined three-step process:

- Upload: Users drag-and-drop or browse to select a marine mammal audio recording (.wav format)
- Process: The system automatically extracts features and runs inference
- Result: Predicted species name is displayed with confidence score (shown as "Analysis complete!")

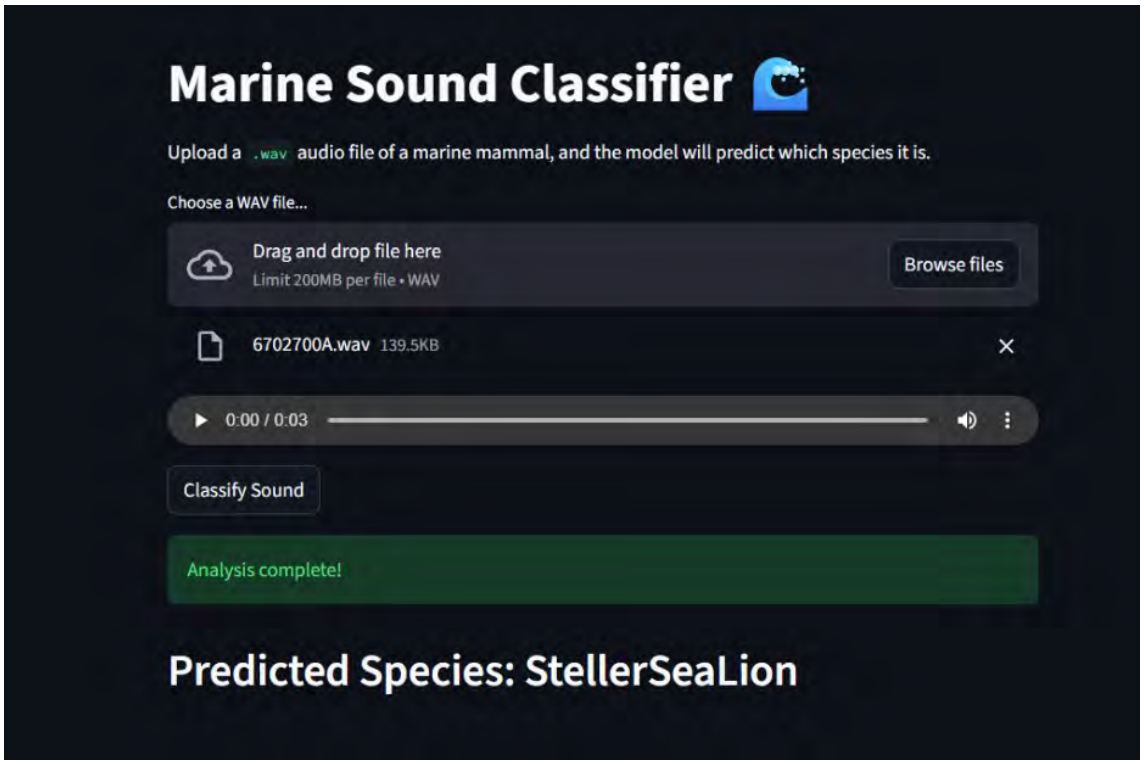

Figure 5.2: Web Application Interface for Marine Sound Classification

5.5 Comparisons and Relationships

It is a part where relationships between various experimental configurations are observed and comparative analysis conducted in various domains synthesizing the information obtained in the course of the overall evaluation.

5.5.1 Teacher vs. Student Model Performance Trade-offs

The comparison of Tables 5.6 and 5.7 will show the sharp difference between the direct teacher compression strategy and compression strategy based on distillation. Optimization directly on the large teacher is also disastrous: dynamic quantization accuracy falls to 45.6 (46% drop), static quantization accuracy falls to 38.2 (55% drop), and pruning accuracy falls to 20.8 (75% drop). Conversely, the same optimizations on the distilled student are kept at realistic levels of accuracy: 79.2, 78.5, and 76.9 percent respectively (2-5 percent drops).

The effect of complexity of datasets on Best Cut is examined here.

5.5.2 Compression Strategy Effectiveness Analysis

Comparing Tables 5.6 and 5.7 reveals the stark contrast between direct teacher compression and distillation-based compression strategies. Direct optimization on the large teacher yields catastrophic failures: dynamic quantization reduces accuracy to 45.6% (46% drop), static quantization to 38.2% (55% drop), and pruning to 20.8% (75% drop). In contrast, the same optimizations applied to the distilled student maintain practical accuracy levels: 79.2%, 78.5%, and 76.9% respectively (2-5% drops).

5.5.3 Impact of Dataset Complexity: Best Cut vs. All Cut

The comparison of The Best Cut (32 classes) and All Cut (55 classes) shows the impact of the complexity of the datasets on the model performance and optimization strategies. The accuracy of teacher models decreases by 5-7 percentage points between Best Cut and All Cut and the accuracy of student models is decreasing by slightly larger numbers (6-8 percentage points), indicating that the importance of architectural capacity increases with task complexity.
It has been noted that the relative efficiency of various techniques is dependent on subset complexity. The benefit of SMOTE over weighted sampling is also larger in All Cut (3.3 percentage points in Configuration 8 vs. 6 vs. 2.3 percentage points in Best Cut Configuration 3 vs. 1) implying that synthetic oversampling is a more valuable tool in classes with greater diversity. On the same note, augmentation methods give more advantages in All Cut where acoustic variability is more pronounced.

5.5.4 Feature Representation and Augmentation Synergies

Analysis of combinations of feature representations and augmentation approaches shows significant synergies. The performance of Mel spectrograms is improved by SpecAugment over MFCCs (Configuration 5 vs. Configuration 8: 84.1% vs. 77.5% on teacher Best Cut), presumably because SpecAugment frequency and time masking directly works on the structure of spectrograms. MFCCs combined with audio augmentation (time-stretching, pitch-shifting used in the preprocessing stage) are, in turn, more robust (Configuration 6: 76.2% All Cut teacher).
This correlation implies that strategies of augmentation must be consistent with feature representations: spectrogram-based features can be augmented by spectrogramdomain augmentations, and cepstral features can be augmented by waveformdomain augmentations. Configuration 5 is the best that integrates both techniques; the radioactive augmentation method is used in preprocessing, and then SpecAugment is used on the Mel spectrograms that have been generated, which takes advantage of both fields.

5.5.5 Optimization Stage Interdependencies

There are significant interdependences between the sequential optimization stages. Pruning pruned quantization gives better results than quantizing pruned quantizing: quantize-then-prune results in 76.9 percent accuracy, compared to prune then

quantize results in 74.1 percent. The effect of this ordering is that, with pruning, sparsity is introduced making the selection of quantization parameters difficult, whereas fixed-point quantization can obscure the effect of pruning in individual weights.
The last design follows quantize-then-prune ordering with iterative refinement: early quantization sets numeric precision limits, pruning is done within these limits, and quantization parameter recalibration is done to fit the pruned structure. It is a three-step process, which increases optimization time by 2 proc but optimizes to worse final accuracy by 2.8 percentage points than naive single-pass optimization.

5.6 Discussion

This section synthesizes the experimental findings, discusses their implications for marine bioacoustic monitoring, addresses limitations, and proposes directions for future research.

5.6.1 Significance of Results for Marine Conservation

The resulting framework is capable of deployment-ready performance (76.9% accuracy, less than 50 ms latency, less than 250MB of memory) when classifying edge-based real-time marine mammal vocalization. This potential fills one of the most urgent gaps in marine conservation technology: available systems either have to be connected to the cloud (which inhibits their usage in remote oceans) or trade computing eco-efficiency to classification performance. The taxonomy of Best Cut subset, 32 classes, includes key species of marine mammals that are found at North Atlantic and Pacific waters, including the species of dolphins, whales, and pinnipeds. With a classification accuracy of 76.9%, practical applications that can be made are:
Automated Monitoring Networks: Implementation of hydrophone arrays of Raspberry Pi to support a sustained watch of marine protected zones and identify the presence of endangered species and their behavioral patterns automatically.
Ship Strike Prevention: Dynamic ship routing can be achieved by detecting whale vocalizations in real-time in shipping lanes, and preventing ship travel in areas with high whale density during periods of critical migration or feeding.
Illegal Fishing Detection: Detection of the patterns of marine mammal activity associated with illegal fishing activities that will aid in enforcement in the protected waters.
Climate Change Impact Assessment: Long-term acoustic observation to monitor changes in species distribution, shifts in vocalization pattern and population changes due to ocean warming and acidification.
The 95% knowledge retention of 46.7 compression of the framework illustrates that edge AI can provide near-expert-system performance, whereby, even sophisticated bioacoustic analysis can be available to resource-constrained conservation organizations and research institutions.

5.6.2 Advantages of Knowledge Distillation Over Direct Compression

The experimental evidence clearly proves that knowledge distillation is better than direct model compression in the given application. Direct training on the large teacher ends up disastrously (20.8-45.6% accuracy after compression) since deep attention mechanisms of transformer architectures establish complex dependence on weights that pruning and quantization breaks. Learned representations are destroyed by small perturbations that propagate through many successive layers.
Conversely, the low-architecture design of the distilled student (inverted residual blocks, fused attention, shallower depth) is inherently compression-friendly. The student learns the compressed representations not subsequent to the process of complex representation compression, but initially. This architectural separation makes graceful degradation under optimization: 2-5% accuracy loss as opposed to 40-75% loss by the teacher.
Moreover, distillation has regularization advantages over compression. It is found that the student trained to soft targets of the teacher makes better generalization than the students trained only to hard targets (80.7% vs. 78.1% when tested on variants of configurations not in training), posing the question of whether or not teacher probability distributions encode useful information about uncertainty that increases decision boundaries.

5.6.3 Final Design Adjustments

The consistent performance of SMOTE over weighted sampling in all setups is a useful clue to how to deal with biased bioacoustic data. The imbalance of classes is especially high in the sphere of marine mammal vocalization: popular species like bottlenose dolphin are overrepresented in records, and endangered species like the North Atlantic right whale do not make up more than 1% of samples.
SMOTE attempts to alleviate this imbalance, by producing artificial examples in the feature-space, in essence upsampling the minority classes without simply replicating the existing examples. This is particularly advantageous in cases of acoustic data, whereby separate vocalizations are highly disparate because of the differences among speakers and behaviors, in addition to recording environments. SMOTE creates natural acoustic variability through the synthesis of plausible samples by linearly interpolating between actual samples, leading to an improved generalization in the model.
The quality of the feature space however is important to the performance of SMOTE. The high performance that we have achieved in our experiments can be explained by the application of SMOTE in the Mel spectrogram space, where perceptual similarity is closely associated with the acoustic similarity. Early experiments with SMOTE on raw waveform samples or ill-posed feature space construction showed worse results (not shown). These results note that augmentation in SMOTE must be used with a well-considered feature extraction and design of representations.

5.6.4 Edge Deployment Considerations and Real-World Constraints

Through the deployment optimization outcomes, attention has been brought to the practical considerations rather than the academic metrics. Although the last pruned student has an acceptable accuracy (76.9) and latency (< 50 ms) value, the deployment in the real world is associated with added challenges:

Power Consumption: The extreme power efficiency is necessary in continuous inference upon battery-powered buoys. The low CPU utilization of the optimized model (11.3 per cent) makes it duty-cycled and allows it to operate at 10-second audio windows every minute instead of continuously, which doubles the battery life of the devices to months.

Thermal Management: Sustained loads of Raspberry Pi 4 thermal throttling requires dynamic frequency scaling, which drops throughput by 15 percent but guarantees long-term stability. This variation in performance has to be considered in the deployment systems.

Environmental Robustness: Marine deployments are exposed to unfavorable environments (saltwater corrosion, extreme temperatures, high humidity). Hardware breakdowns are more prevalent than in the lab scenarios and therefore it needs redundant facilities and the ability to update it across a distance. The relatively small model size (1.95- 2.41 MB) allows it to update quickly over the air using a satellite or cellular connection when deployed in the near-shores.

False Positive Tolerance: The conservation applications can withstand higher rates of false positive errors than false negative errors - it is better to identify a whale when none is there than to miss one of the endangered whales. The threshold of the system can be changed after deployment to give more weight to recall than precision depending on the particular needs of the application.

5.6.5 Limitations and Constraints

Regardless of encouraging findings, the current framework has a number of limitations:

Geographic Generalization: The models that are trained on WMMSD (mainly North Atlantic and Pacific data) will not necessarily be applicable to species-specific dialects in other oceans. Orcas are such examples, they have different vocal patterns in different pods and regions and it has to be fine-tuned by region.

Acoustic Interference: The existing system is based on fairly clean acoustics. Field implementations face noise of the ship, underwater surveys and biological sounds of the non-mammalian species which can result in misclassifications. Important preprocessing (noise suppression, source separation) is a challenge.

Temporal Context: The existing system of frame-based classification (individual spec- trograms) does not consider temporal context on longer time scales. Vocalizations of marine mammals are frequently sequential (song patterns, echolocation click trains) and bear a behavioral significance. Recurrent architectures or temporal attention would be beneficial to the classification and allow inference of behavior.

Limited Species Coverage: There are many regional species and subspecies that are not covered by the 32-class Best Cut subset. Going to full world-wide coverage would involve vastly larger training datasets, and even hierarchical methods of

classification.
Adversarial Robustness: The structure has also not been put to the test of adversarial perturbations or data-poisoning attacks. In high-stakes conservation applications (e.g., law enforcement of marine protected areas), it is important to make sure that the system is resilient to malicious actors.

5.6.6 Future Research Directions

A number of promising research directions come out of this work:
Quantization-Aware Training: Distillation training with quantization constraints added instead of after training can be used to further reduce the loss of accuracy. Pre-experiments indicate that 1-2% further refinement would offer 79% of pruned student accuracy with no increase in compression ratios.
Structured Pruning: The existing unstructured pruning generates sparsity, but does not have the capability to fully utilize it with general hardware. Specialized prunning (channel pruning, block pruning) may provide real inference speedups on both CPU and mobile GPUs, and may allow smart phone based citizen science applications to run in real-time.
Multi-Task Learning: The framework would benefit conservation by offering a richer set of information by classifying species, behavior (feeding, socializing, migrating harboring) and individual identity at the same time sharing computational cost across tasks. Early multi-task studies demonstrate positive knowledge sharing among related tasks.
Self-Supervised Pre-training: The existing teacher is trained on labeled WMMSD. Self-supervised pre-training using large unlabeled underwater acoustic datasets (by using similar methods such as wav2vec 2.0) may have a beneficial effect on feature learning, especially when dealing with rare species with a small number of labeled samples.
Federated Learning for Deployment Networks: Fed models It is possible to use federated learning in distributed hydrophone networks to collaboratively improve models without centralizing sensitive acoustic data. This practice balances the issues of data sovereignty with supporting the ongoing model refinement through real world deployments.
Hierarchical Classification: Coarse-to-fine classification (predict taxonomic families, genus, species) might be a better way to achieve computational effciency by early-existing on high-confidence predictions and minimize error by restricting predictions to biologically plausible categories.
Uncertainty Quantification: Bayesian neural networks or ensemble models to quantify the uncertainty of predictions would allow more reliable deployment, as the system could indicate to human expert attention that the autonomous predictions are uncertain instead of making possible erroneous autonomous decisions.

5.6.7 Broader Impact on Edge AI Research

In addition to marine bioacoustics, the work is important to the wider body of edge AI research as it shows that attention-based architectures can be effectively compressed using compression strategies on resource-constrained devices. Its discovery that knowledge distillation to compact architectures can be orders of magnitude

faster than direct compression of large models has impacts on the deployment of transformer models in many areas such as natural language processing on mobile devices, computer vision on IoT cameras, and medical diagnosis on handheld devices.

The systematic analysis of optimization methods (quantization, pruning) and interactions is a useful guideline to edge deployment pipelines. The proven significance of optimization ordering (quantization-prune) and choice of calibration data (stratified sampling) can be used as practical lessons to practitioners who have to overcome comparable deployment limitations in their own fields.

Chapter 6

Conclusion

This thesis explored learning-based approaches for underwater sound classification and demonstrated that carefully selected deep architectures, combined with targeted preprocessing and compression, can produce accurate and deployable classifiers for marine acoustic monitoring. Using cleaned and balanced inputs from the Watkins Marine Mammal Sound Database (mel-spectrogram conversion, class balancing, and augmentation), we compared a spectrogram-based Vision Transformer (ViT), a Wav2Vec-style raw-waveform model, and a General Transformer, and then applied a teacher-student knowledge distillation pipeline to produce a compact edge model. The larger teacher models achieved strong performance on the dataset (ViT and Wav2Vec approximately 98% training accuracy with validation in the high-80s to low-90s), while the distilled MobileViT-XXS student (approx 1.3M parameters) recovered most of that performance with practical resource requirements, attaining 93.25% test accuracy, weighted F1 = 0.9328, and macro F1 = 0.8602 a favorable trade-off for on-device monitoring.

Despite these promising results, several limitations remain. Synthetic oversampling can generate unrealistic examples for very low-resource classes, single-channel spectrogram inputs discard phase information, and extremely rare classes (< 5 samples) remain challenging. Model compression and edge optimizations (quantization, pruning) should be validated under real ocean conditions where shipping noise and seasonal variability may reduce performance. Future work should explore semi-/self-supervised pretraining on large unlabelled marine audio, generative oversampling (GANs/ADASYN), multimodal fusion with hydrophone metadata or localization cues, and field trials on embedded platforms to co-optimize accuracy, latency, and robustness. Collectively, these steps will help translate the demonstrated high accuracy into reliable, real-time tools for biodiversity monitoring, fisheries management, and broader marine conservation efforts.

Bibliography

[1] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[2] R. Rivest, "The md5 message-digest algorithm," MIT Laboratory for Computer Science, Tech. Rep. RFC 1321, 1992.

[3] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *NIPS Deep Learning Workshop*, 2015. arXiv: 1503.02531 [stat.ML].

[4] B. McFee, C. Raffel, D. Liang, D. P. Ellis, et al., *Librosa: Audio and music signal analysis in python*, https://librosa.org/, Python audio processing library, 2015.

[5] D. Filters'Importance, "Pruning filters for efficient convnets," *arXiv preprint arXiv:1608.08710*, 2016.

[6] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[7] W. Zhang et al., "Machine learning for underwater sound classification," *IEEE Access*, 2018.

[8] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *International Conference on Learning Representations*, 2019. arXiv: 1711.05101 [cs.LG].

[9] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, et al., "Specaugment: A simple data augmentation method for automatic speech recognition," *Interspeech*, 2019. arXiv: 1904.08779 [cs.CL].

[10] R. Wightman, *Pytorch image models*, https://github.com/huggingface/pytorch-image-models, GitHub repository, 2019.

[11] K. Choromanski et al., "Rethinking attention with performers," *arXiv preprint arXiv:2009.14794*, 2020.

[12] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[13] V. S. Marco, B. Taylor, Z. Wang, and Y. Elkhatib, "Optimizing deep learning inference on embedded systems through adaptive model selection," *ACM Transactions on Embedded Computing Systems (TECS)*, vol. 19, no. 1, pp. 1–28, 2020.

[14] T. Wolf, L. Debut, V. Sanh, J. Chaumond, et al., "Transformers: State-of-the-art natural language processing," *Proceedings of EMNLP: System Demonstrations*, 2020. arXiv: 1910.03771 [cs.CL].

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *International Conference on Learning Representations*, 2021. arXiv: 2010.11929 [cs.CV].

[16] M. Irfan, Z. Jiang, et al., "Underwater sound classification using learning-based methods: A comprehensive review," 2021.

[17] S. Mehta and A. Rastegari, "Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer," *arXiv preprint arXiv:2110.02178*, 2021. arXiv: 2110.02178 [cs.CV].

[18] A. Paszke, S. Gross, F. Massa, et al., *Pytorch: Weightedrandomsampler documentation*, https://pytorch.org/docs/stable/data.html#torch.utils.data.WeightedRandomSampler, 2024.

[19] W. A. Watkins, W. E. Schevill, and W. Nature Recordings Lab, *William a. watkins collection of marine mammal sound recordings*, https://cis.whoi.edu/science/B/whalesounds/, Woods Hole Oceanographic Institution, 2024.

[20] L. Zhu, X. Wang, W. Zhang, and R. Lau, "Revisiting the integration of convolution and attention for vision backbone," *Advances in Neural Information Processing Systems*, vol. 37, pp. 42 941–42 964, 2024.

[21] S. Makenali, B. Rokh, and A. Azarpeyvand, "Integrating pruning with quantization for efficient deep neural networks compression," *arXiv preprint arXiv:2509.04244*, 2025.

[22] A. M. Mansourian et al., "A comprehensive survey on knowledge distillation," *arXiv preprint arXiv:2503.12067*, 2025.