MULTIMODAL APPROACH FOR SEAFLOOR IMAGE CLASSIFICATION USING FEATURE-LEVEL DATA FUSION

by

Kedar Pandya

Submitted in partial fulfillment of the requirements
for the degree of Master of Computer Science

at

Dalhousie University
Halifax, Nova Scotia
April 2025

This thesis is dedicated to my mother and father.

Table of Contents

Abstract

Mapping seafloor substrates is crucial for understanding benthic habitats and monitoring their changes. Traditionally, this task involves manually annotating underwater images, an approach increasingly impractical due to growing data volumes. This thesis investigates a multimodal deep-learning approach for classifying seafloor substrates by integrating visual images with sonar data (backscatter and bathymetry) collected from the Bay of Fundy. To identify the role of wider spatial context, two sonar data sampling methods were compared: single-point sampling, which assigns a single sonar value based on image coordinates, and context-based sampling, which incorporates sonar data from adjacent measurements. Experiments demonstrated that context-based sampling, which provided a wider spatial context, significantly improved substrate classification accuracy.

Feature-level data fusion strategies were then evaluated by encoding sonar data using multilayer perceptrons and extracting marginal representations from different layers of the sonar encoder. These representations were fused with visual features extracted via convolutional neural networks. The effective fusion strategy, established through these experiments, was subsequently implemented using a pre-trained Vision Transformer and ResNet50 models as image encoders. While ResNet50-based multimodal models showed only moderate improvements relative to a baseline trained solely on visual data, ViT-based fusion models achieved significantly larger gains, exceeding two standard deviations and improving accuracy by approximately 35% for challenging substrate types with substantial class overlap.

Overall, this research demonstrates that multimodal learning frameworks can effectively leverage complementary features from sonar data sampled to provide a wider spatial context and visual data, resulting in substantial improvements in seafloor substrate classification accuracy. These findings establish a robust foundation for further research in multimodal oceanography and benthic mapping.

Acknowledgements

This thesis results from hard work, inspiration, and support from many wonderful people besides myself. I want to extend my foremost gratitude to Dr. Thomas Trappenberg. Over the past two years, he has been much more than a supervisor; he has been a mentor, a trusted friend, and a source of encouragement. His confidence in me pushed me to go beyond my limits, and his thoughtful feedback kept me grounded and focused. I also want to recognize the Department of Oceanography at Dalhousie University and the SEAM Lab for providing the dataset for my research. I am also sincerely grateful to Isaac Xu for his continuous feedback and guidance during the often overwhelming research journey. His encouragement helped shape the inquisitive mindset that you see reflected throughout this work. My thanks also go to Dr. Martin Gilles, whose insights into the latest industry trends and research methods have enriched my approach. In addition, I appreciate all the members of the HAL Lab for our many inspiring discussions, which have sparked creativity and provided much-needed motivation.

I want to give a big hug to my parents for their constant and silent support throughout this journey. I am grateful to them for providing me with the opportunity to study at one of the most prestigious universities in Canada. Their belief in me has been a driving factor in my attainment of success and in showing my best work. I also express deep gratitude to my dear friend and roommate, Ritvik Voleti. He has been my rock through challenging times, always ready to lift my spirits and remind me of the bright side. I'm also thankful to Meet, Nihal, Yaksh, and Vraj, who have been more like family than friends, always there with care, understanding, and support. Lastly, I want to share my deepest appreciation with the person who means the world to me, *mi Lila*. Without her constant support and motivation, I would not have been able to overcome the many obstacles along the way. Her presence has been a steady beacon of hope and encouragement, and I am forever grateful.

Chapter 1

Introduction

1.1 Motivation

The benthic (seafloor) environment is important to ocean health and global economies. Understanding its rapidly changing dynamics helps capture the impacts of oceanic climate change and guide protective policies. Benthic mapping involves measuring, analyzing, and classifying the seafloor substrates (composition), topography, and biota (fauna). Oceanographers typically create these maps using imagery, which is then annotated manually. However, the increasing data volume has exceeded human annotation capacity, leading to the adoption of machine learning algorithms [1][2][3][4]. In addition to image data, sonar data is also used by oceanographers for benthic studies. Sonar provides complementary information about the seafloor that cannot be captured through visual imagery alone. For example, cameras record visual features like colour and texture, while sonar systems capture depth and substrate characteristics over larger areas. We hypothesize that by integrating sonar data as an additional modality, machine learning models could better classify benthic substrate images, addressing the limitations of models trained on image data alone.

In this work, multimodal learning is applied to classify the benthic substrate using underwater images and sonar data collected using multibeam sonar from the Bay of Fundy. Multimodal learning is a concept in deep learning that integrates different data modalities like visual or acoustic data to analyze the subject, enabling the model to process information with additional context. The sonar data used in this work has two types: backscatter (which gives information about the seafloor substrates) and bathymetry (which gives information about the seafloor's depth) measures. In this research, we use feature-level fusion [5], where features from each modality are learned by independent networks and combined before classification. This work aims to answer the following research question:

Does the addition of sonar modality improve the classification capabilities of the deep learning model for benthic substrate images in the Bay of Fundy dataset? And if so, what is the best approach?

To answer this main research question, we identified three sub-research questions and investigated them in detail:

1. Does using a wider spatial context for the sonar data help with the benthic substrate classification?

To create a structured multimodal dataset from the visual and sonar data, we needed to determine if sampling sonar data with a wider spatial context benefits the classification of substrates. In Subsection 3.1.2, two data sampling methods were proposed for sonar data: single-point sampling and context-based sampling. In single-point sampling, an area of 50 x 50 m^2 is sampled, whereas Context-based sampling covers a wider area of 150 x 150 m^2. However, this increased coverage might cover multiple substrate classes and create noise in the data. Thus, we experimented (detailed in Subsection 3.3.1 with results presented in Section 4.1) with data sampled from both of these methods using a Random Forest model to classify the substrate images and concluded that context-based sampling was more effective for classifying benthic substrate images.

2. Which feature-level fusion framework is the most effective for multimodal seafloor substrate classification?

To answer this question, we designed experiments with variations of dual and triple encoder frameworks. Both frameworks used a CNN as the image encoder and MLP networks as the sonar encoders. In the dual-encoder framework, the visual features were fused with either backscatter or bathymetry data, while the triple-encoder framework fused visual features with both sonar data types, each encoded by separate MLPs. Variations in fusion were introduced by extracting feature representations from the first and last layers of the sonar encoders and fusing them with the image features. All models were trained from scratch, and the variations focused exclusively on the sonar encoders to identify the most effective feature-level fusion strategy for

the classification problem. These experimental details are further elaborated in Subsection 3.3.2, with results provided in Section 4.2. Our experiments showed that using first-layer representations in the dual-encoder framework and last-layer representations in the triple-encoder framework yielded the best classification accuracies, making these frameworks the basis for further experiments.

3. Can the effective multimodal framework generalize well to pre-trained image networks?

To address this question, we replaced the CNN with pre-trained ViT (Vision Transformer) and ResNet50 models trained on BenthicNet [6], a global dataset of benthic images. The same dual and triple-encoder architectures were used, with the dual-encoder relying on first-layer sonar representations and the triple-encoder on last-layer representations, as established from the previous sub-research question's results. Detailed experimental procedures are provided in Subsection 3.3.3, and the results are discussed in Section 4.3. Incorporating sonar data significantly improved the ViT models' classification accuracy, demonstrating the proposed framework's effectiveness and addressing the main research question.

The remainder of this thesis is structured as follows: Chapter 2 provides the background, covering vocabulary development and dataset discussion; Chapter 3 outlines the methodology, detailing dataset pre-processing, the experimental setup, and evaluation methods; Chapter 4 and Chapter 5 present the results, discussions, and conclusions derived from this work's experiments.

1.2 Related Works

Many studies have explored multimodal learning in fields such as computer vision[7], remote sensing[8], and healthcare[9], yet only a limited number of investigations focus on ocean science.

Notably, Wang et al. [10] introduced an Ocean Data Fusion Network (ODF-Net), demonstrating how multimodal learning can effectively combine routine observations, satellite imagery, and bathymetry. Their network employs an intermediate fusion strategy: each modality is encoded separately and then fused by leveraging spatial attention mechanisms to prioritize the most significant features. Their work, which led to the 3D sea temperature ODF-ST dataset, showed marked improvements in creating

more accurate and consistent ocean data. Unlike their focus on generative tasks with satellite and bathymetric data, our work utilizes multimodal learning with benthic images and sonar data, specifically evaluating models for classification performance.

Rao et al.[11] presented a feature-level fusion model with denoising autoencoders to learn a generative relationship between underwater images and bathymetric data. The work of Rao et al. inspires our work and builds on their research by supporting the visual modality with backscatter data in addition to the bathymetric data. Moreover, we also utilize different fusion models and frameworks to investigate the effect of multimodal learning on benthic substrate classification in the Bay of Fundy.

Chapter 2

Background

2.1 SONAR Systems

Sound Navigation and Ranging, or sonar, is used in ocean exploration to help scientists create underwater maps. There are two main types of sonar: active and passive. Passive sonar system [12][13] does not emit a signal but works by receiving signals. They use a special microphone called hydrophones that only listen to sounds coming from the water. On the other hand, active sonar[12] is a system where the device sends sound waves created by electrical signals in the transducer and receives them through receivers. These returning signals are then turned into sonar images, which help create different types of underwater maps. Researchers widely use active sonar systems to create benthic maps[13][14]. There are various types of active sonar systems that scientists use based on the requirements of their research. In this study, the sonar data was collected using Multibeam Echosounder sonar (MBES).

MBES, also known as multibeam sonar [15], was invented in the 1970s by Renard et al. [16]. A typical MBES system comprises an array of transducers, a user interface for controlling the system, and supplementary equipment such as GPS, altitude sensors, and sound-velocity sensors. The MBES transducers are mounted on the bottom of a boat or ship and emit sound wave swaths directly beneath the vessel. Unlike single-beam systems emitting a single sound pulse, MBES transmits multiple beams across a wide swath simultaneously. Each ping produces many overlapping beams that create a dense grid of measurements, capturing data with finer spatial detail and higher overall resolution. Importantly, in most MBES systems, the multiple beams are transmitted at the same frequency during each ping. Beamforming techniques then steer these beams in different directions, maximizing coverage without changing frequency mid-ping. Some advanced systems can operate at multiple frequencies but typically do so in separate pings rather than simultaneously mixing frequencies. MBES is a popular choice for oceanographers engaged in benthic mapping [17]. MBES

systems collect two data types, namely bathymetry and backscatter.

Bathymetric data provides information about the depth of the seabed. Sonar systems capture this data by calculating the time difference between sound signals' emission and reception, adjusting for factors such as water temperature, salinity, and pressure. This data allows scientists to create detailed 3D maps of the ocean floor, revealing features such as underwater mountains, valleys, and plains.

On the other hand, backscatter data reveals the seafloor's substrate information and physical characteristics. Oceanographers collect this data by measuring the intensity of the returned sound signal. Different materials on the ocean floor, such as sand, rocks, or mud, reflect sound in unique ways. For instance, hard surfaces like rocks send a stronger signal, while softer surfaces like mud reflect weaker ones. This difference in intensities helps scientists categorize the seafloor according to the substrates found. Backscatter data assists in identifying underwater hazards or objects, such as shipwrecks, which consist of materials differing from those on the natural ocean floor. In this work, we utilized both these types of sonar data types for experimentation.

2.2 Dataset

The dataset used in this thesis is collected from the Bay of Fundy. It is located in the North Atlantic Ocean near the coasts of Nova Scotia and New Brunswick in Canada and Maine in the United States. This dataset is a composite of two distinct modalities: sonar and visual. The sonar data includes backscatter and bathymetry information of the region and is structured in a raster format but unlabeled, while the visual dataset is both structured and labelled. To use these datasets, we were required to pre-process them (shown in Section 3.1) and create a structured multimodal dataset for further experimentation.

2.2.1 Sonar Data: Backscatter and Bathymetry

Todd et al.[18][19] gathered the sonar backscatter and bathymetry data from the Bay of Fundy between 1992 and 2009. Since these datasets were collected during different surveys, they had to be merged to create a comprehensive backscatter and bathymetry dataset for the entire bay. However, combining the data from multiple surveys is

challenging due to the rapid dynamics of benthic environments and variations in data collection and processing methods across surveys. To fix this problem, Misiuk et al. [20] proposed a Bulk-Shift method for harmonizing the sonar dataset, which was used by Haar et al.[21] to harmonize the sonar data from the Bay of fundy region.

The Bulk-Shift method uses the overlapping area between different data sets to perform relative statistical calibration. It aims to find the common area between two datasets and then create a model to estimate the error based on the backscatter or bathymetry values of the set to be shifted. By treating the error as the response variable, the Bulk-Shift method allows for visualization of the location and magnitude of the error concerning the dataset being shifted. In their work, Haar et al.[21] tested several models for fitting and predicting the error, including simple and multiple linear regression, and found that these models effectively reduced discrepancies between datasets. The authors employed a multi-step approach, prioritizing the combination of the most similar and spatially extensive datasets first, followed by the integration of the most dissimilar legacy datasets. This strategy allowed them to generate harmonized sonar backscatter and bathymetry data in their study area.

The backscatter and bathymetry data are stored in raster images. In this format, each pixel stores the strength of the sonar wave as it bounces off the ocean floor for the backscatter data. On the other hand, for the bathymetry data, each pixel represents the measurement of the seafloor's depth from the water level. The backscatter data is 5005x4893 pixels, and the bathymetry data is 3982x4581 pixels in size. Additionally, this data format stores information in layers or bands. The first band of the raster contains the raster values for backscatter or bathymetry, whereas the second band contains the geographical coordinate values of the respective raster data. These coordinates are essential for linking the sonar and visual modalities to create a multimodal dataset.

2.2.2 Visual Data

The Bay of Fundy images used in this work were gathered by Wilson et al.[22] to form detailed benthic maps of the region. These images are a subset of BenthicNet[6], a larger dataset of underwater images taken from different places around the world.

The images collected between 2017 and 2019 include 1,879 unique pictures taken

from 281 sites. Figure 2.1 shows the geospatial raster image of backscatter and bathymetry data overlaid with these sites. Chartered vessels collected the pictures using drop-down video drifts with high-quality 4K cameras. These cameras drifted passively for 5 to 15 minutes at 1-2 meters above the seafloor. Each drift covered distances between 100 and 500 meters. Wilson's team systematically chose pictures from the recorded video showing the seafloor. They selected ten pictures at 10-meter intervals for each drift from every 100-meter video segment. The geographical coordinates of these pictures were also recorded while these pictures were captured with a location accuracy within ± 16 meters.

Wilson et al.[22] annotated these images using benthoscape types (the spatial distribution and composition of benthic habitats on the seafloor). Lowe et al. [6] later converted these annotations into the Collaborative and Automated Tools for Analysis of Marine Imagery (CATAMI) system [23] within BenthicNet to ensure consistent labelling across different subsets in BenthicNet. This scheme is designed to classify visual images in a hierarchical context and is used to label the ocean floor's physical and living parts. Figure 2.2 shows the CATAMI annotation hierarchy used in the Bay of Fundy subset of the BenthicNet dataset.

The CATAMI labelling scheme is both hierarchical and multi-label in nature. In this work, we perform classification independently at each CATAMI hierarchical level, treating these levels as separate multi-label classification tasks rather than attempting to perform hierarchical classification across levels. At a given depth, the combination of labels assigned to an instance is referred to as a multi-label set or a pattern of labels, and the set of all possible label combinations at that depth is known as the multi-label space.

An analysis of the dataset reveals that it is imbalanced across all CATAMI hierarchical levels, as shown in Figure 2.3. At hierarchical level 1, the Unconsolidated (soft) class (brown bar) and instances with multiple labels dominate (purple bar), while the Consolidated (hard) class (teal bar) is underrepresented. At hierarchical level 2, the imbalance shifts toward prominent multi-label classes such as Cobbles-Pebble/gravel (blue bar) and Pebble/gravel-Sand/mud (green bar). Additionally, the classification becomes complicated due to a larger multi-label space at Level 2, which makes it a crucial depth for model evaluation in our experiments. hierarchical level

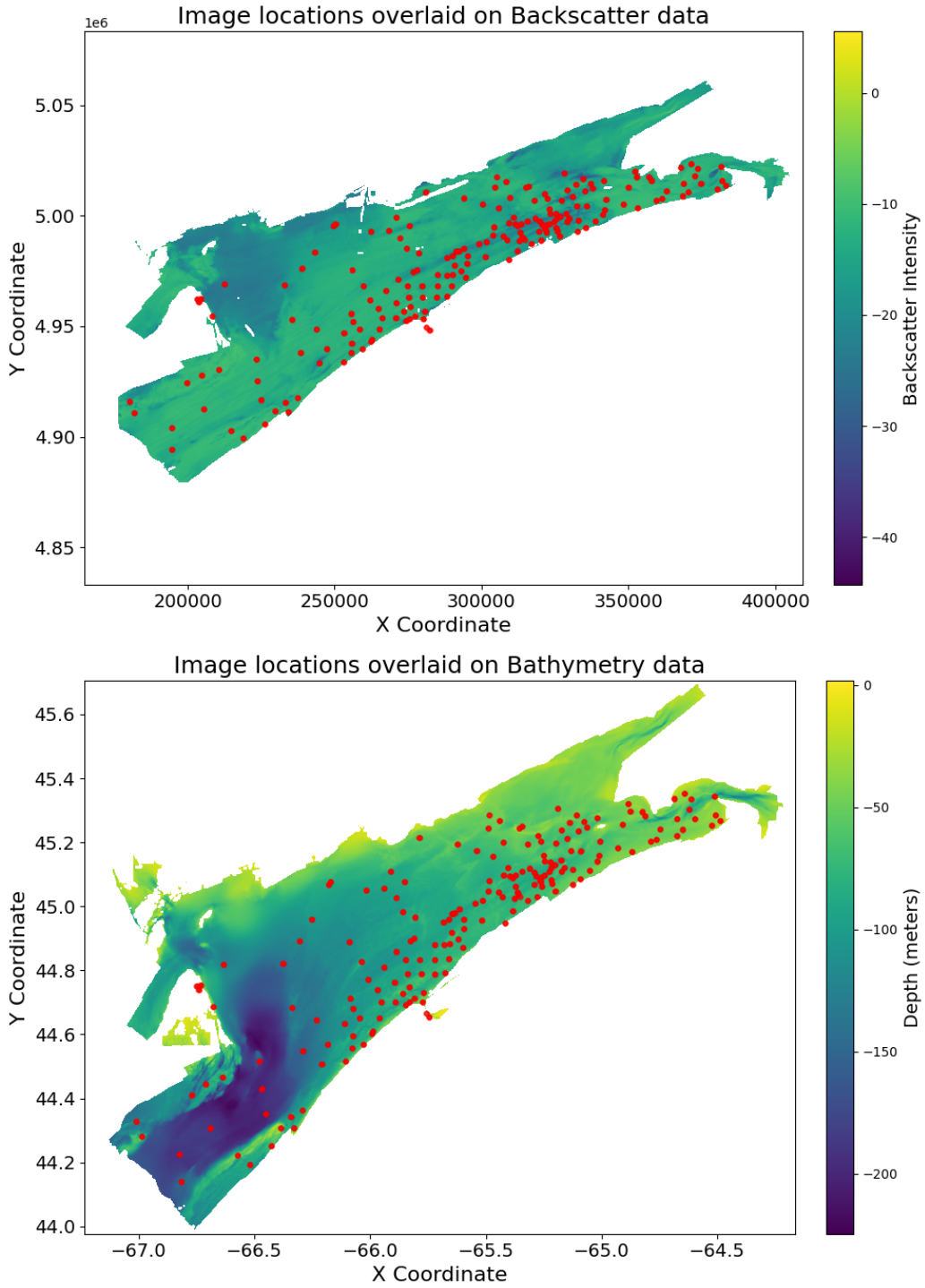

Figure 2.1: The First plot is a geospatial intensity plot of backscatter data, and the second plot corresponds to the bathymetry data. The red dots on the plot indicate the points in the Bay of Fundy from where the underwater images were taken. The darker regions in the backscatter plot imply low-intensity regions, while the lighter regions represent high-intensity areas. Similarly, the darker region in the bathymetry plot indicates a deeper seafloor, whereas the lighter region suggests a shallower one.

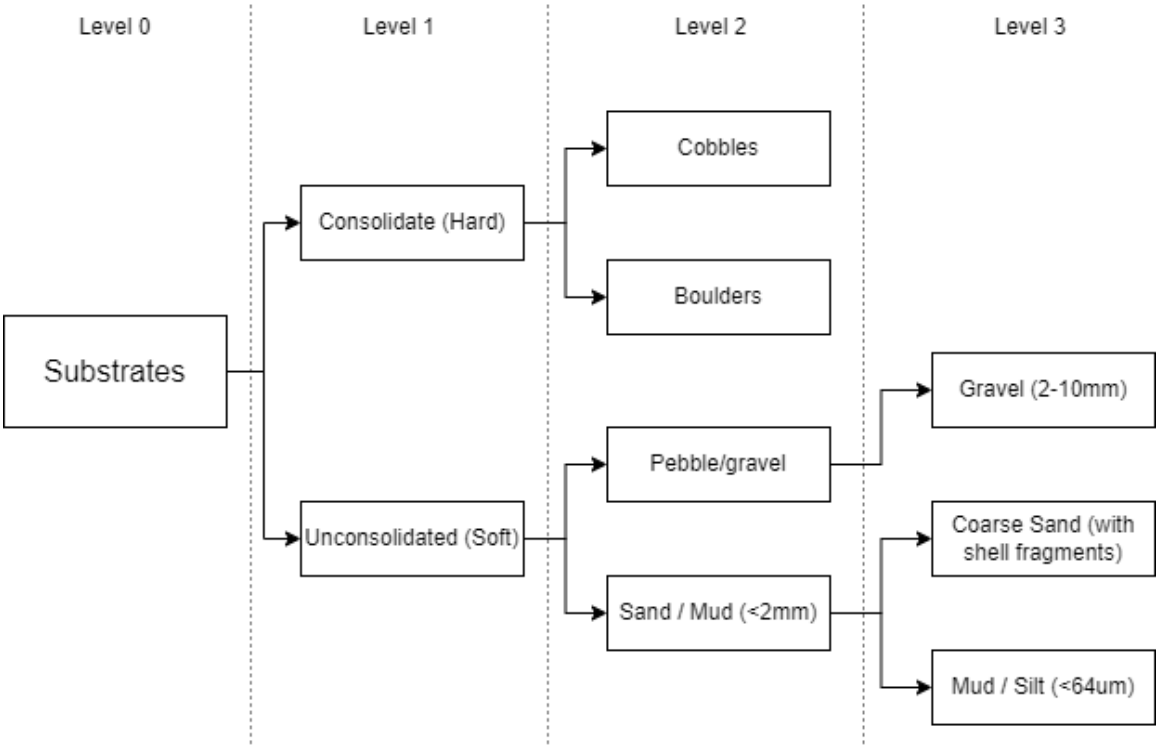

Figure 2.2: The CATAMI hierarchy used for labelling Bay of Fundy images, illustrating the classification framework for marine substrates. The hierarchy progresses from general substrate categories at Level 0 to increasingly detailed divisions, culminating in specific substrate types at Level 3.

3 exhibits the most severe imbalance, with the Mud/silt class (pink bar) having the fewest examples by a significant margin.

Moreover, at Level 3, the Coarse sand-Gravel multi-label class (orange bar) always co-occurs, with no instances of these two labels appearing independently. This perfect co-occurrence simplifies classification at this depth, effectively reducing it to a binary classification problem, as the model only needs to predict the presence of one of the co-occurring label groups. However, this co-occurrence is only observed in this dataset and might not be a trend in other benthic substrate datasets.

2.3 Pre-trained BenthicNet Models

This work uses pre-trained deep learning models, specifically ResNet-50 and Vision Transformer (ViT). Xu et al. [24] pre-trained these models on the unlabeled portion of the BenthicNet dataset. We fine-tuned these pre-trained models on our multimodal dataset, which is a subset of the labelled BenthicNet dataset. This dataset is entirely

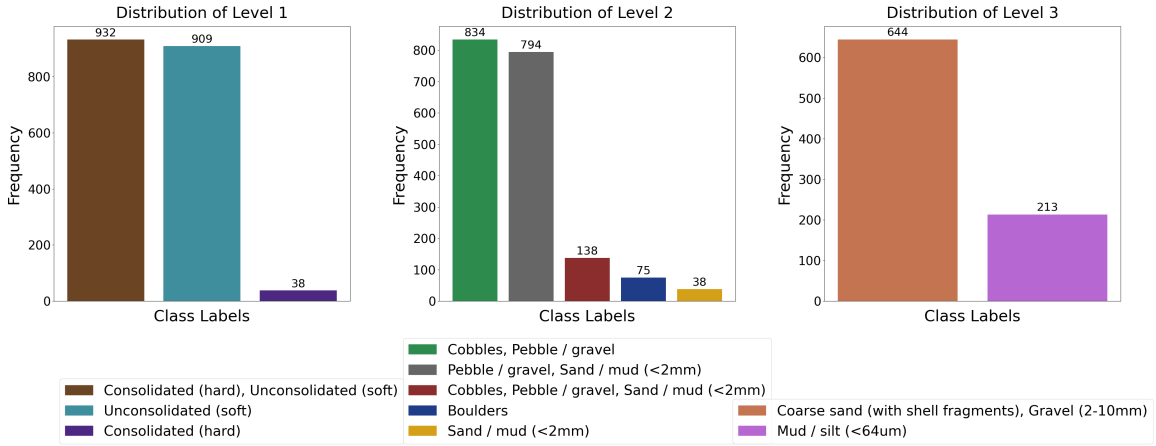

Figure 2.3: Class frequency distributions at different levels of the CATAMI hierarchy. The plots represent the frequencies of unique multi-label patterns, highlighting the imbalanced distribution of classes across Levels 1, 2, and 3.

separate from the unlabeled data used for pre-training, ensuring no data leakage occurs.

2.3.1 ResNet50

ResNet50 [25] is a deep convolutional neural network architecture that introduces residual learning through skip connections, which help mitigate model degradation issues in deep networks. The network consists of 50 layers organized into residual blocks, allowing it to learn residual functions based on layer inputs. This design significantly improved performance on image classification tasks and remains a widely used model in computer vision.

Barlow Twins [26] is a self-supervised learning (SSL) framework that was used to pre-train the ResNet50 model used in this work. This framework maximizes the similarity between augmented views of the same image while reducing redundancy between different components of their representations. Two augmented views of each image are processed through the same ResNet50 encoder to produce embeddings, which are used to compute a cross-correlation matrix. The loss function encourages diagonal elements of the matrix (representing similarity between the same features) to approach 1, while pushing off-diagonal elements (representing redundancy) toward 0. This simple yet effective approach has shown competitive performance compared to other self-supervised methods.

2.3.2 Vision Transformer (ViT)

A Vision Transformer (ViT) [27] leverages transformer-based models, originally developed for text processing, to perform visual data-related tasks. It works by splitting an image into fixed-size patches (e.g., 16x16 pixels), flattening these patches into vectors, and processing them as a sequence with positional embeddings. Unlike convolutional neural networks (CNNs), which process images using sliding windows and pooling operations, ViT treats image patches as tokens and uses self-attention mechanisms to capture relationships between patches for image understanding.

MoCo v3 [28] is a self-supervised learning (SSL) framework that was used to pre-train the ViT model in this work. It employs a momentum contrast approach with an asymmetric architecture consisting of two encoders: an online encoder (the primary ViT being trained) and a momentum encoder. The momentum encoder maintains a moving average of the online encoder to stabilize training. During pre-training, two augmented views of the same image are created. One view is processed by the online encoder, while the other is processed by the momentum encoder. The training objective is to maximize the similarity between the representations of both views (positive pairs) while pushing apart representations of different images, using a contrastive loss. MoCo v3 further enhances training by removing projection heads and leveraging block-wise self-attention, making it well-suited for transformer-based models.

2.4 Multimodal Data Fusion

Multimodal data fusion combines information from different sources (data modalities) to give the model more context for the downstream task [29][30]. There are various fusion techniques used to merge distinct modalities. One of those methods is element-wise operations, where the feature vector of different modalities are reshaped to the same size and combined using addition or multiplication [31]. Another technique, called bilinear pooling, takes the outer product of the feature vectors to capture complex interactions between features from different sources [32][33]. However, the most popular method is concatenation due to its execution simplicity and minimal computational requirement compared to other methods [34][35][36][37]. There are

different frameworks for the deep learning models to fuse data modalities at various processing stages. These frameworks are: 1) raw fusion (early fusion), 2) feature-level fusion (intermediate fusion), and 3) decision-level fusion (late fusion) [34][31][30].

Raw fusion combines inputs from different sources into a single vector before feeding it into a machine-learning model. This framework is simpler to implement but learns the relationship between the modalities at a very low-level [31]. In decision-level or late fusion, different sources combine their data by merging the decisions made by individual models[5][31][34]. In this approach, each data modality is processed separately by its model, which learns to make predictions. The final decision for the downstream task is typically made by combining the outputs of these models, such as class probabilities or decision scores, either equally or with weighted averaging. While this allows each modality to have a strong feature representation, it fails to capture the correlation between modalities [31]. Intermediate fusion addresses the drawbacks of both raw and decision-level fusion by effectively learning the relationships between modalities.

In Intermediate Fusion, the fusion of modalities occurs between the partially learned feature vectors, thus giving it a name: feature-level fusion[5][31]. Here, each modality is passed through a separate encoder network, as shown in Figure 2.4, and learns an independent feature representation of the modality, which is referred to as marginal representation. These representations reflect the network's ability to extract standalone features before fusion operations[5][38] [34]. After learning these marginal representations, the system merges them into a joint representation that captures how the different modalities relate. This combined representation can then be used for tasks such as classification or regression.

Intermediate fusion methods have several advantages over early and late fusion approaches. In intermediate fusion, the system first learns patterns from each type of modality before combining them [5]. This characteristic is helpful when working with data from different sources having unique traits, like formats or temporal resolutions [39]. Intermediate fusion also allows more control over selecting marginal representations from the network layers for data fusion, helping the system understand how the various data formats connect [40]. Unlike late fusion, which only combines the decisions from each data modality, intermediate fusion can learn how the features from

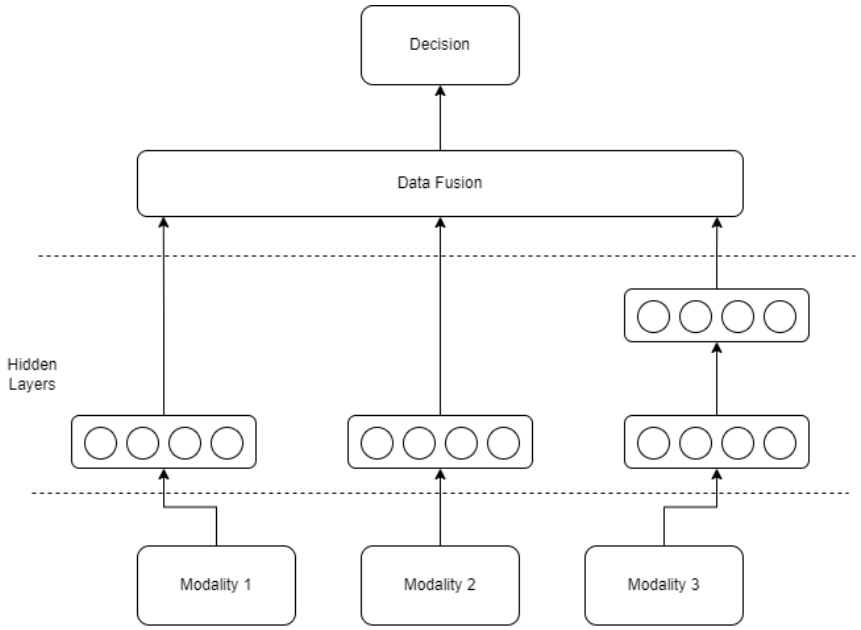

Figure 2.4: Illustration of the intermediate feature-level fusion process, where representations from multiple modalities (Modality 1, Modality 2, and Modality 3) are combined at a shared fusion layer before making a decision. The diagram highlights the flow of information and integration of features across modalities.

each data modality correlate. Additionally, intermediate fusion makes it easier to see how each data modality affects the result since the model processes each separately, which is harder to do in late fusion. Due to these advantages over other frameworks, we employed the feature-level fusion framework for multimodal learning in this work. This framework was first implemented using basic two-layer CNN image encoders for the visual modality and then using the pre-trained image models as image encoders.

Chapter 3

Methods

3.1 Data Pre-processing

As mentioned in Section 2.2, the dataset used in this work was derived from two distinct modalities: sonar data and visual data. We required the integration of sonar and visual data to create a structured multimodal dataset. Figure 3.1 illustrates the data pre-processing pipeline used in this work to create the multimodal dataset.

3.1.1 Data Integration

The data integration phase began with extracting the raster values from the first and coordinate information from the second band of the sonar raster, illustrated as the first step in Figure 3.1. As the image data was already structured, acquiring the coordinate values did not require any process. These geographical coordinates of images and sonar are calibrated using the Coordinate Reference System(CRS), a generalized way to define locations on Earth's surface. The CRS used for the bathymetry and image data was EPSG:4326, while the CRS of the backscatter dataset was EPSG:32620. To address the difference in the coordinate reference systems (CRS) between the backscatter data and the images and bathymetry data, the Python libraries PyProj [41] and RasterIO [42] were used to reproject the image data to match the CRS of the backscatter data. After aligning the datasets through careful reprojection and coordinate matching, the next crucial step, before integrating the data, was to sample the sonar data in a manner that aligns with the spatial scale of the benthic images.

3.1.2 Sonar Data Sampling

Previous studies have explored different approaches to sampling sonar data for substrate classification. For instance, Gregrid et al.[43] sampled a larger area of 100 x 100 m^2 for bathymetry data, which enhanced substrate classification in deeper benthic

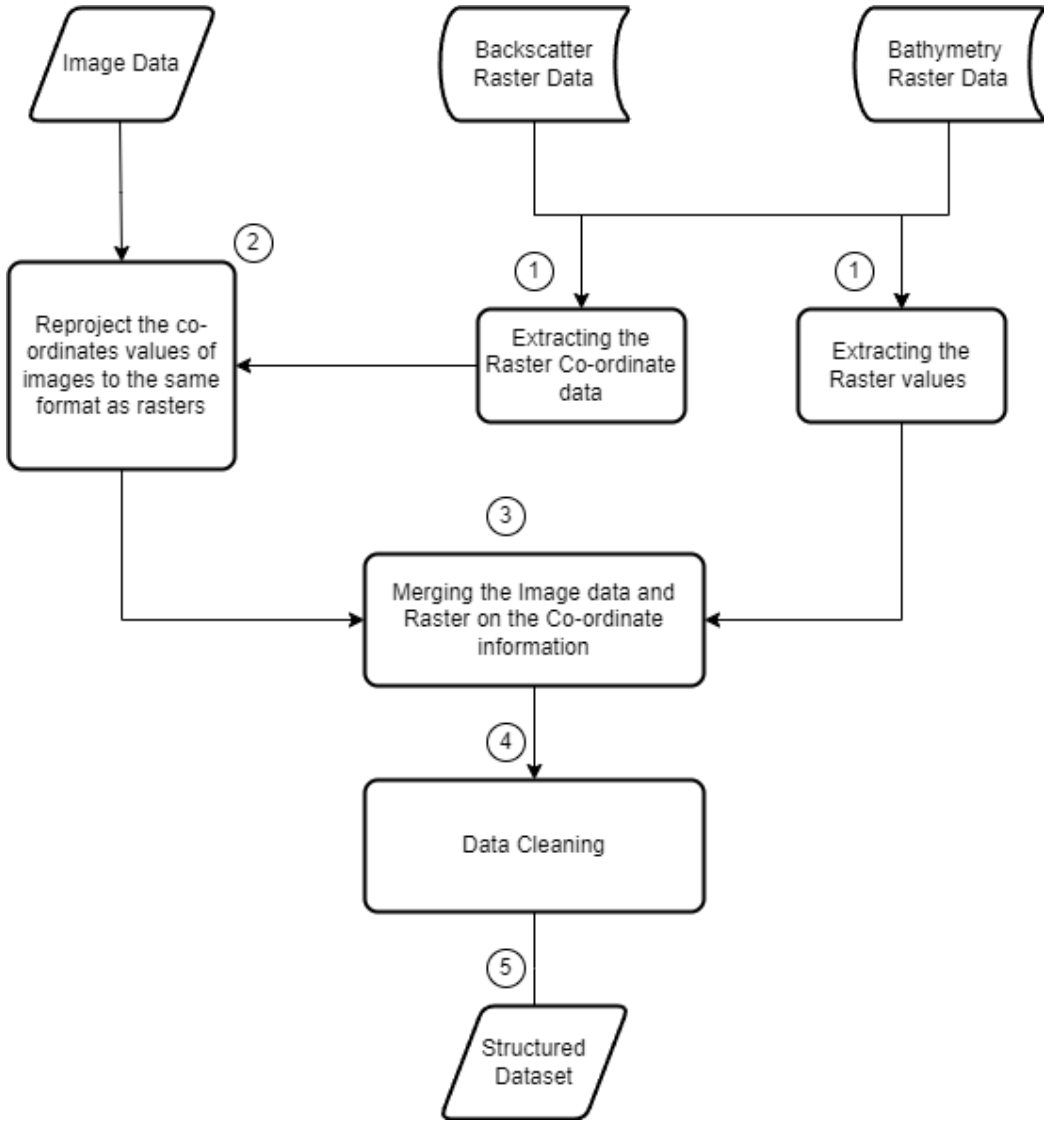

Figure 3.1: The unstructured backscatter and bathymetry data first go through an extraction process in which raster values and coordinates are gathered. The images are then mapped to these raster values using the coordinate values of images and raster pixels. This resultant dataset then goes through a data cleaning process to finally obtain a structured multimodal dataset.

environments (greater than 50m depth) compared to the data sampled from a 20 x 20 m^2 area, which was more effective for coastal regions. Similarly, Rao et al.[11] used a grid patch of 15 x 15 raster pixels, which covered a 22.4 x22.4 m^2 area for bathymetry sampling.

In the Bay of Fundy dataset used in this study, the geospatial coverage of a single raster pixel for backscatter and bathymetry data is 50 x 50 m^2. The seafloor depths of the Bay of Fundy go up to 240 meters [44], which might make this resolution not sufficient to classify the benthic substrates according to the study conducted by Gregrid et al. [43]. To address this, two sampling methods were implemented to map the sonar data to the benthic images, one covering a smaller area and the other incorporating a larger region.

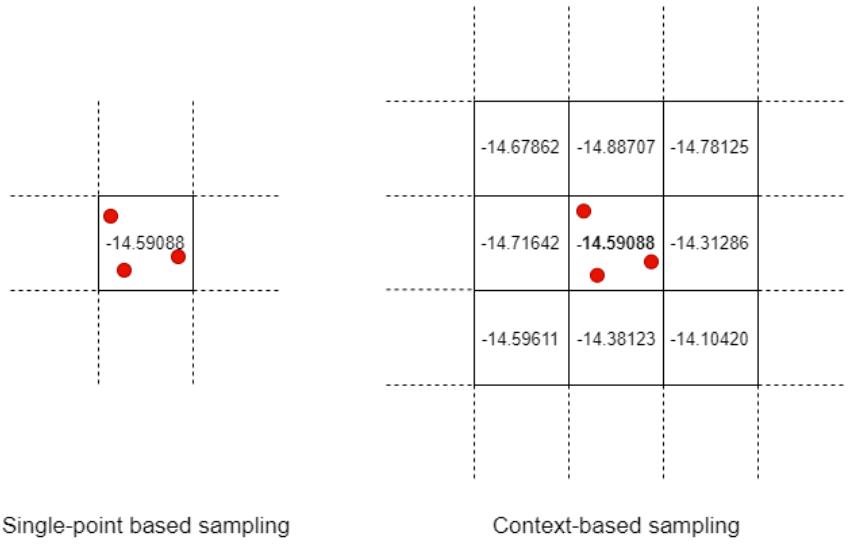

Figure 3.2: Visualization of data sampling methods: Single-point sampling assigns a single value (e.g., backscatter or bathymetry) to each image (red dots), while context-based sampling incorporates the neighbouring pixel values around each image.

The first method, called single-point sampling, assigns each image a single backscatter and bathymetry value based on its coordinates. This method provides a spatial coverage of 50 x 50 m^2 per image. Figure 3.2 illustrates this approach. However, the images cover much smaller spatial areas as they were captured 1-2 meters above the seafloor. As a result, multiple images often fall within the bounds of a single raster pixel, giving similar raster values to multiple images.

The second method, context-based sampling, expands on single-point sampling by incorporating the values of neighbouring raster pixels. A 3 x 3 patch of raster pixels

is captured for each image, where the central pixel corresponds to the raster pixel containing the image. The adjacent pixels surrounding this central pixel are also included, linking the image to multiple backscatter and bathymetry values (rather than just one). This approach increases the spatial coverage to 150 x 150 m^2. Figure 3.2 illustrates this method as well. Like the single-point sampling method, the same values can be assigned to multiple images.

Both sampling methods were evaluated to determine their effectiveness, addressing the first sub-research question of this study. Experiments were conducted to analyze how the spatial coverage differences between the two methods impacted the classification of benthic substrates

Data sampled from both of these sampling methods were then integrated with the image data based on their geographic coordinate information between sonar and image data. This integration allowed the annotations from the image dataset (based on the CATAMI scheme) to be transferred to the sonar. Figure 3.3 illustrates an example of the structured data generated after merging the image and sonar datasets. After merging the sampled sonar data with the corresponding image coordinates, the consolidated dataset was refined through data cleaning processes in the ultimate step of data pre-processing, as shown in Figure 3.1.

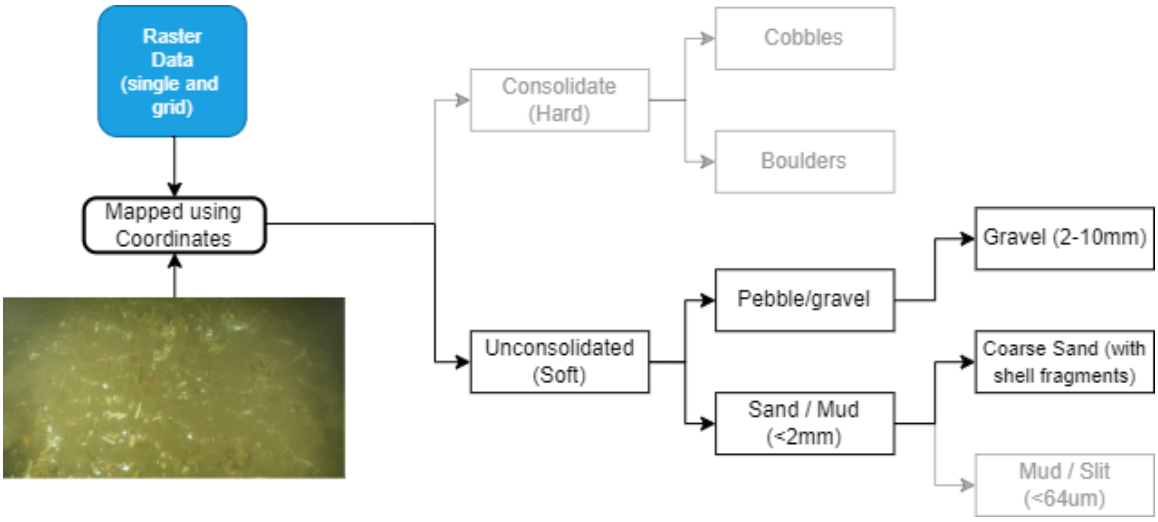

Figure 3.3: Example of mapping raster data (sonar) to visual data from the Bay of Fundy. The image's CRS is reprojected to match the raster CRS, allowing CATAMI multi-label annotations from the image to serve as ground truth for the corresponding raster data. This method supports mapping using both sampling methods: Single-point and context-based raster pixels.

3.1.3 Data Cleaning and Imbalance Analysis

As part of a conventional data cleaning check, we verified that no duplicate entries were introduced during the integration of the raw sonar and image data. Additionally, we identified null entries in the dataset. Since the dataset followed a hierarchical structure based on the CATAMI scheme, it contained higher-level classes with fewer samples. When these hierarchical levels were separated into individual subsets, null entries were created. To address this issue and ensure the dataset was consistent and ready for further processing, we removed the null entries.

After removing null entries, the backscatter and bathymetry data values were normalized using the StandardScaler function from the Scikit-Learn library. Additionally, all images were rescaled to dimensions of 224 x 224 x 3 using the transform function from PyTorch, ensuring uniformity in image resolution. After completing these preprocessing steps, the final structured dataset comprised a combination of multiple data types, including imagery and acoustic data obtained from sonar, prepared for downstream analysis.

3.2 Training Configurations

3.2.1 Dataset split

To obtain a test set for final evaluation from the new structured multimodal dataset, we divided it into two distinct sets: an 80% training set and a 20% test set. This initial split was performed using a fixed seed (42) and maintained class distribution via the stratify parameter from Scikit-Learn.

CATAMI Hierarchy	Number of Training Samples	Number of Testing Samples
Level 1	1503	376
Level 2	1503	376
Level 3	685	172

Table 3.1: Number of Data instances in the training and test set across the three CATAMI hierarchy levels.

The training set was then further partitioned into training and validation subsets in an 80:20 ratio for the purpose of cross-validation, again using stratification to ensure

balanced classes. We employed a 5-fold cross-validation procedure on this training data, and the models developed during this process were ultimately evaluated on the separate test set. This workflow ensures that the test set remains untouched during model tuning and is used solely to assess the generalizability and variance of the final models. Table 3.1 shows the number of samples in the train and test set for each CATAMI hierarchical level.

3.2.2 Model Blocks

To construct the multimodal models in Experiments 2 (Subsection 3.3.2) and 3 (Subsection 3.3.3), two fundamental components, which are Convolutional Neural Network (CNN) and a Multilayer Perceptron (MLP) were used. Figure 3.4 illustrates the underlying layers used in these building blocks.

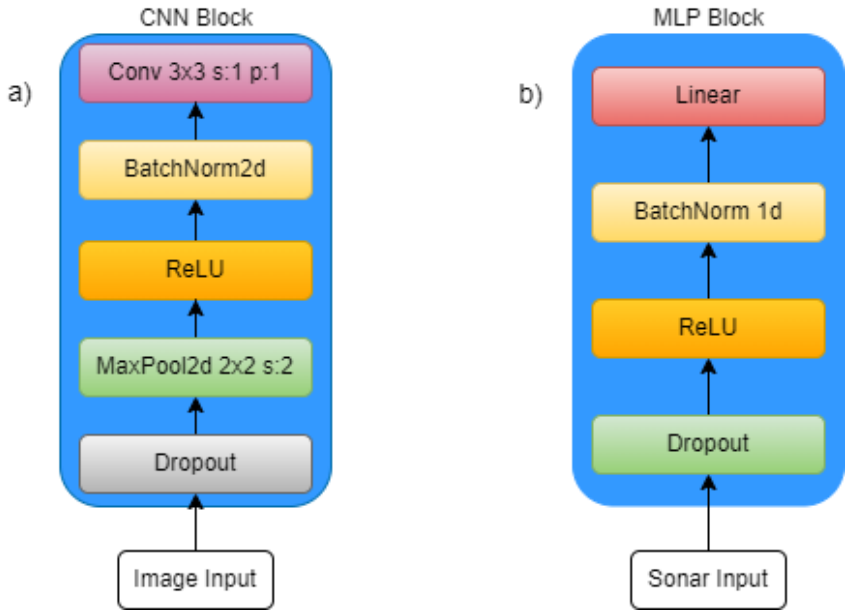

Figure 3.4: Building blocks used in the multi-modal model. The CNN block (a) processes visual data, while the MLP block (b) processes sonar input

The CNN block comprises a convolutional layer with a kernel size 3x3, with stride and padding set to 1. A batch normalization layer follows this convolutional layer. The feature vector of the image then passes through the ReLU activation layer before going into the Maxpool layer. The final layer is an optional dropout layer, which reduces the number of learnable parameters, further reducing overfitting. The MLP block uses a similar structure, but instead of a convolutional layer, it has a fully

connected linear layer and doesn't require a Maxpool layer. In addition to these network blocks, a projection layer was also used. Projection layers are known to act as information bottlenecks, encouraging the network to learn robust, generalized representations while reducing the risk of overfitting [45][46]. These projection layers were a single-layer MLP followed by a BatchNorm and ReLU. During experimentation, these network blocks were used to create the sonar and image encoders.

3.2.3 Loss Function, Optimizer and Epochs

To train the models used in this work, we used the Binary Cross-Entropy with the Logit Loss function. This loss function treats each label as a separate entity to be predicted and computes the loss independently for each label in the label space. This method puts no constraint on selecting only one label and explicitly allows multiple labels to be predicted, which is the requirement of this work. The function is defined as

$$\text{BCEWithLogitsLoss}(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \cdot \log(\sigma(\hat{y}_i)) + (1 - y_i) \cdot \log(1 - \sigma(\hat{y}_i))]$$

where N is the batch size, y_i is the true label, $\hat{y}_i$ is the logit, and $\sigma(\hat{y}_i) = \frac{1}{1+e^{-\hat{y}_i}}$is the sigmoid function.

The loss function outputs logits rather than probabilities, requiring an explicit sigmoid activation to convert logits into probabilities. We used a default threshold of 0.5 to obtain binary predictions. Additionally, we employed the Adam optimizer, commonly used in classification tasks. For experimentation, we trained all the neural networks for 50 epochs using the specified loss function and optimizer.

3.2.4 Regularization Techniques

In this work, we used regularization techniques of dropout and weight decay. Additionally, we apply drop path rate, also known as stochastic depth [47], to both ViT and ResNet-50 architectures, used in Experiment 4 (Subsection 3.3.4). Unlike dropout, which operates at the neuron level, stochastic depth randomly drops entire blocks. The drop probability increases linearly with depth, meaning deeper layers are more likely to be dropped than earlier ones. In Vision Transformers (ViTs), this

means dropping blocks containing multi-head self-attention and feed-forward networks. However, in ResNet architectures, stochastic depth functions differently. Instead of dropping transformer blocks, it stochastically bypasses entire residual blocks during training. This regularization has proven effective in deeper networks to improve generalization capacities [27][48].

3.2.5 Hyperparameter Tuning

The Optuna[49] framework was used to tune the hyperparameters for all the models in this work. This framework dynamically samples a unique set of hyperparameters in the search space using a sampler strategy. We used the default sampler strategy, which works by classifying the parameters into promising and unpromising sets and then drawing parameters from the promising set, increasing the likelihood that future trials will yield better results[50]. We created the objective function of the Random Forest models to maximize validation accuracy, setting the number of trials to 1000. In contrast, for deep learning networks, we defined the objective function to minimize validation loss, a standard practice in hyperparameter pruning for such models. In this case, the number of trials was limited to 100, comprising 10 training epochs. The hyperparameters identified during this tuning phase were subsequently employed to train and cross-validate the models.

3.3 Experimental Setup

3.3.1 Experiment 1: Spatial Feature Analysis: Single vs Context-based Data Sampling

This experiment aimed to answer the first sub-research question, which is to examine the impact of a wider spatial context for benthic substrate classification. We evaluated the performance of machine learning models trained solely on sonar data using two different data sampling methods, mentioned previously in Subsection 3.1.2.

Baseline Setup

The baseline model used in this experiment is a Bayesian classifier, with kernel density estimates. Unlike the machine learning models, the entire dataset was provided

as input to this baseline model without splitting it into training and testing sets. This approach was chosen because the baseline model aims to provide near-optimal estimated scores for the method, serving as a point of comparison for evaluating the performance of the machine learning models. For simplicity, only the single-point values of the dataset were used in this model, as incorporating context-based features would have significantly increased the data's dimensionality.

The model used data distributions obtained via kernel density estimates of the single-point values as input. These distributions were visualized using the kdeplot function from Python's Seaborn library, which smooths the frequency values using a Gaussian function. This smoothing makes the plots analogous to histograms while providing an estimate of the actual density function of the data. One of the

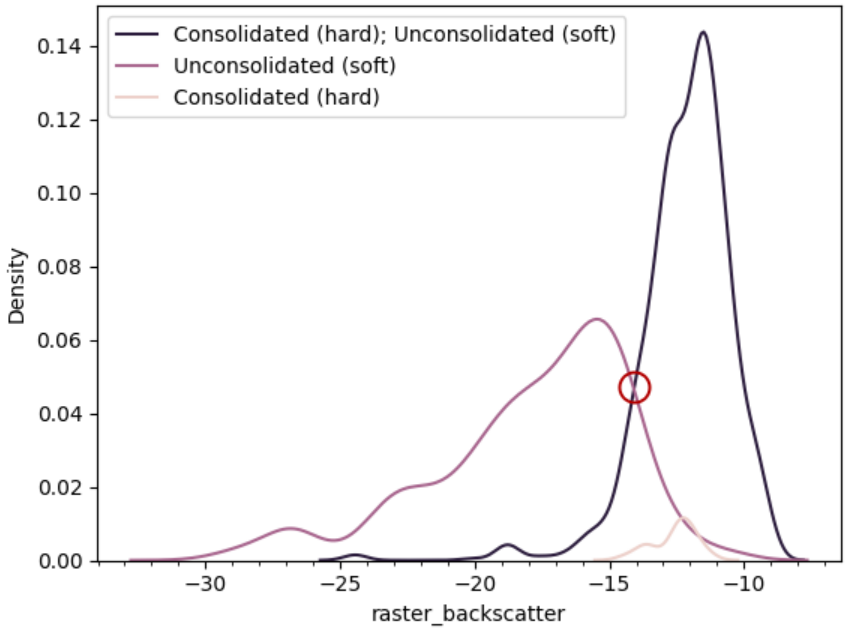

Figure 3.5: Kernel Density plot of CATAMI hierarchical level 1 label for backscatter dataset. The red circle indicates the decision point decided by the BayesianKDE model to classify between the Unconsolidated(soft) class and the Consolidated(hard)-Unconsolidated(soft) multi-label class. Here, the model is not able to distinguish the Consolidated (hard) class due to its distribution being completely overlapped.

density distributions used for the BayesianKDE is illustrated in Figure 3.5. Specifically, this distribution shows CATAMI hierarchical level 1 for the backscatter dataset. The model was given six different density distributions to analyze two data formats (backscatter and bathymetry) and three different levels.

To find the separation between different classes, the model looks for points where the estimated density values of classes are equal. These points are called decision boundaries; the relationship is written as:

$$f_i(x) = f_j(x)$$

Here, $f(x)$ is density of the data and i, j represents the different classes existing in the dataset.

From the Figure 3.5, we can observe that the density of Unconsolidated (soft) and multi-label pattern of Consolidated (hard)-Unconsolidated (soft) classes become equal at approximately -14.086. Therefore, this point becomes the decision boundary of this data. Any values before this point are predicted to be from the Unconsolidated (soft) class, and values after this point are assigned to the Consolidated (hard)-Unconsolidated (soft) class by the BayesianKDE model. Here, the Consolidated (hard) density distribution lies completely inside the multi-label class's density distribution, resulting in no decision boundary being identified for the class and creating a bias in the model.

Experimental Setup for Random Forest Models

The machine learning model selected for this experiment is Random Forests, chosen for several reasons. First, backscatter and bathymetry features have lower dimensionality than image data, making their processing more straightforward. Additionally, tree-based models, like Random Forests, classify data by selecting decision points based on the information gained from the features. While Random Forests use this information-driven decision-making, the BayesianKDE model relies on kernel density estimates, resulting in a key difference in determining decision points. However, the classification approach remains similar, as both models rely on data-driven methods to identify decision points. This similarity means that performance differences between these two models can be more confidently attributed to the effects of the spatial context (i.e., single-point versus context-based inputs) rather than fundamentally different classification strategies. Furthermore, Random Forest models are widely used by oceanographers for tasks involving backscatter or bathymetry data [51] [52] [53]. We trained the Random Forest model using both single-point-based inputs and

CATAMI Hierarchy	Input Format	Criterion	Number of Estimators	Max Depth	Min Sample Split	Min Sample Leaf	Max Features
Backscatter							
Level 1	Single	Entropy	10	40	10	8	None
	Grid	Entropy	10	30	2	8	sqrt
Level 2	Single	Entropy	10	10	20	4	log2
	Grid	Entropy	50	20	20	2	sqrt
Level 3	Single	Gini	200	40	5	4	sqrt
	Grid	Gini	100	20	2	1	sqrt
Bathymetry							
Level 1	Single	Gini	10	30	20	2	sqrt
	Grid	Entropy	50	10	10	2	None
Level 2	Single	Gini	10	10	20	1	sqrt
	Grid	Entropy	200	20	5	1	None
Level 3	Single	Gini	10	40	5	8	sqrt
	Grid	Entropy	10	30	2	8	sqrt

Table 3.2: Hyperparemeter values obtained by Optuna for the Random Forest models using backscatter and bathymetry data.

context-based inputs. The tuned hyperparameters for the Random Forest models used in this experiment are provided in Table 3.2.

3.3.2 Experiment 2: Dual Encoder Data Fusion Analysis

This experiment addresses the second sub-research question by exploring a dual-encoder model framework that combines visual and sonar modalities through separate encoders. Only one data type from the sonar modality, either backscatter or bathymetry, was used for training. Based on the results of spatial feature analysis experiments, data sampled using the context-based approach was selected for this study [a]. In this experiment, we tested several simple model variations to evaluate different fusion points within the dual-encoder architecture.

[a] A detailed discussion of these results is provided in Section 4.1

Experimental setup for Dual-Encoder fusion

The dual-encoder model architecture effectively processes two types of data using both sonar and image encoders. The sonar encoders utilize a Multi-Layer Perceptron (MLP), while the image encoders consist of a Convolutional Neural Network (CNN) to extract feature representations.

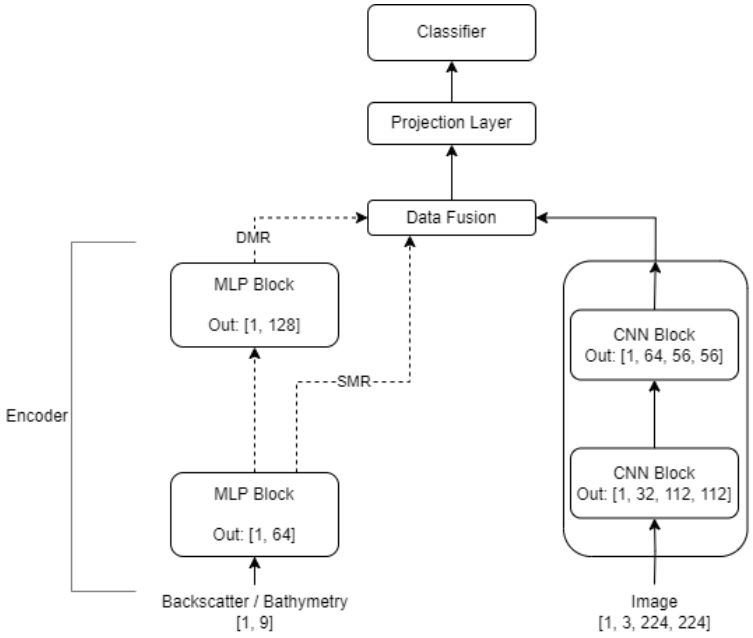

Figure 3.6: Dual-encoder architecture used in this work. Marginal representations of the modalities from varying blocks (layers), shallow (SMR) or deep (DMR), are obtained and fused via a data fusion layer, followed by a projection layer and classifier for the final output. Dual-encoder architecture used in this work. Example feature dimensions are shown at each stage: image tensors are in the format $[b, c, h, w]$ and sonar vectors are in the format $[b, \text{features}]$, where b is the batch size.

Figure 3.6 illustrates the structure of the fusion models, which include both MLP and CNN encoders. Each encoder consists of two blocks (layers) for MLP and CNN, shown in Figure 3.4 (a) and (b). To test the experimental hypothesis, marginal representations [5][31] were extracted from different blocks of the MLP encoder.

Two variations of the multimodal model were developed to evaluate the effectiveness of combining features from different layers. The first model used shallow marginal representations (SMR) from the first block of the MLP encoder, while the second model used deep marginal representations (DMR) from the second block. In

both models, these features were combined through simple concatenation in the fusion layer. The resulting fused feature vector was passed through a projection layer, followed by a classifier. The tuned hyperparameters of both the sub-models are shown in Table 3.3. The learning rate in this experiment was set to $1e-5$.

Model	**Level 1**		**Level 2**		**Level 3**	
	WD	**D/O**	**WD**	**D/O**	**WD**	**D/O**
Backscatter Data (bs)						
SMR	10^{-7}	0.05	10^{-7}	0.05	10^{-7}	0.45
DMR	10^{-5}	0.20	10^{-7}	0.15	10^{-4}	0.80
Bathymetry Data (bm)						
SMR	10^{-6}	0.10	10^{-4}	0.10	10^{-7}	0.30
DMR	10^{-7}	0.05	10^{-4}	0.05	10^{-5}	0.50

Table 3.3: The Optuna tuned Weight Decay (WD) and Dropout (D/O) values for dual-encoder models using Backscatter and Bathymetry Models.

3.3.3 Experiment 3: Triple Encoder Data Fusion Analysis

This experiment expands on Subsection 3.3.2, focusing on the second sub-research question to find an effective fusion framework for a triple-encoder multimodal architecture. In contrast to the dual-encoder method used in Experiment 2, this setup employed three distinct encoders for images, backscatter, and bathymetry data. Different variations of the triple encoder were explored during these experiments to determine an effective fusion strategy.

The multimodal model framework is illustrated in Figure 3.7. Unlike Experiment 2, which used a single MLP encoder, this setup employs two distinct MLP encoders to process backscatter and bathymetry data separately. This design enables independent feature extraction from each sonar data type, enhancing the ability to capture unique characteristics from both. Similar to the earlier experiment, we examined Shallow Marginal Representations (SMR) and Deep Marginal Representations (DMR) from the two MLP encoders, resulting in four unique models.

This setup resulted in the formation of four models, which were:

1. Model using SMR for both the backscatter and bathymetry, called bsSMR-bmSMR

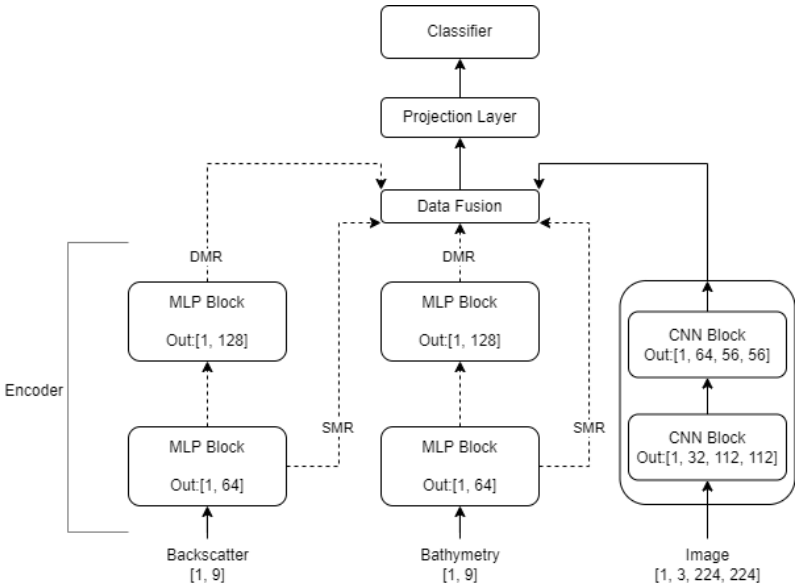

Figure 3.7: Framework of the triple-encoder models used in these experiments. Marginal representations of the sonar data type from varying blocks, shallow (SMR) and deep (DMR), are obtained for fusion with the image representations. Dual-encoder architecture used in this work. Example feature dimensions are shown at each stage: image tensors are in the format $[b, c, h, w]$ and sonar vectors are in the format $[b, \text{features}]$, where b is the batch size.

2. Model using SMR for backscatter and DMR for bathymetry, called bsSMR-bmDMR

3. Model using DMR for backscatter and SMR for bathymetry, called bsDMR-bmSMR

4. Model using DMR for both backscatter and bathymetry. called bsDMR-bmDMR

Table 3.4 shows the fine-tuned hyperparameter values used for the sub-models used in the triple-encoder framework. All the models in this experiment were trained with a learning rate of $1e-5$.

3.3.4 Experiment 4: Feature-Level Fusion Using Pre-Trained Encoders

In this experiment, we evaluate whether the effective multimodal framework (obtained from Experiments 2 and 3) can generalize well to the pre-trained image models. To do this, we utilize image models pre-trained on BenthicNet, specifically Vision Transformer (ViT) and ResNet50, as image encoders for multimodal learning [24].

Model	Level 1		Level 2		Level 3	
	WD	**D/O**	**WD**	**D/O**	**WD**	**D/O**
bsSMR-bmSMR	10^{-5}	0.05	10^{-5}	0.05	10^{-5}	0.10
bsSMR-bmDMR	10^{-7}	0.55	10^{-4}	0.05	10^{-6}	0.05
bsDMR-bmSMR	10^{-7}	0.05	10^{-7}	0.05	10^{-5}	0.15
bsDMR-bmDMR	10^{-5}	0.05	10^{-7}	0.05	10^{-6}	0.05

Table 3.4: The Optuna tuned Weight Decay (WD) and Dropout (D/O) values for triple-encoder models using Backscatter and Bathymetry Models.

Building on the results of Subsection 3.3.2 and Subsection 3.3.3 [b] we evaluate the model's performance in both dual-encoder and triple-encoder frameworks. In the dual-encoder framework, shallow marginal representations of sonar data are fused with image representations extracted from the pre-trained encoders. In the triple-encoder framework, deeper marginal representations of backscatter and bathymetry are combined with the image representations from the pre-trained encoders to assess the impact of richer feature integration.

The baseline models for this experiment were the pre-trained ViT and ResNet50 models fine-tuned on only the visual modality. The learning rate of the baseline model was set to $1e-6$, and the other tuned hyperparameters are shown in Table 3.5.

Model	Level 1		Level 2		Level 3	
	WD	**(D/O, DPR)**	**WD**	**(D/O, DPR)**	**WD**	**(D/O, DPR)**
Backscatter Data (back)						
RN50	10^{-6}	(0.20, 0.30)	10^{-6}	(0.10, 0.10)	10^{-7}	(0.30, 0.05)
ViT	10^{-7}	(0.10, 0.00)	10^{-5}	(0.10, 0.00)	10^{-6}	(0.25, 0.00)

Table 3.5: The Optuna tuned Weight Decay (WD) and Dropout (D/O) and Drop Path Rate (DPR) values for baseline ResNet50 and ViT models.

The dual-encoder framework used in this experiment builds on the design shown in Figure 3.6, with modifications to incorporate pre-trained ResNet50 and ViT models as the image encoders in place of CNN-based encoders. These pre-trained models were fine-tuned by unfreezing their layers and removing their classifier layer to output marginal representations of the images. For the sonar data, a single block of MLP

[b] Detailed results and discussion given in Section 4.2

was employed to generate shallow marginal representations. The fine-tuned hyperparameters for these dual-encoder frameworks are detailed in Table 3.6.

Model	Level 1		Level 2		Level 3	
	WD	**(D/O, DPR)**	**WD**	**(D/O, DPR)**	**WD**	**(D/O, DPR)**
Backscatter Data						
RN50	10^{-4}	(0.25, 0.25)	10^{-6}	(0.10, 0.15)	10^{-6}	(0.05, 0.30)
ViT	10^{-7}	(0.05, 0.05)	10^{-4}	(0.25, 0.00)	10^{-4}	(0.05, 0.00)
Bathymetry Data						
RN50	10^{-6}	(0.25, 0.30)	10^{-6}	(0.05, 0.00)	10^{-5}	(0.05, 0.10)
ViT	10^{-4}	(0.10, 0.00)	10^{-4}	(0.15, 0.00)	10^{-7}	(0.05, 0.00)

Table 3.6: The Optuna tuned Weight Decay (WD) and Dropout (D/O) and Drop Path Rate (DPR) values for ResNet50 and ViT-based dual-encoder models.

The triple-encoder framework used in this experiment is based on the architecture shown in Figure 3.7. The MLP encoders for both backscatter and bathymetry were designed with two MLP blocks each to generate deeper marginal representations. The image encoder in this architecture also utilized pre-trained ResNet50 and ViT models, as described in the dual-encoder setup. The tuned hyperparameters for the triple-encoder models are detailed in Table 3.7.

Model	Level 1		Level 2		Level 3	
	WD	**(D/O, DPR)**	**WD**	**(D/O, DPR)**	**WD**	**(D/O, DPR)**
RN-ALL	10^{-7}	(0.10, 0.10)	10^{-7}	(0.15, 0.15)	10^{-3}	(0.15, 0.25)
ViT-ALL	10^{-6}	(0.10, 0.05)	10^{-4}	(0.10, 0.05)	10^{-6}	(0.05, 0.00)

Table 3.7: The Optuna tuned Weight Decay (WD) and Dropout (D/O) and Drop Path Rate (DPR) values for ResNet50 and ViT-based triple-encoder models.

The learning rate configuration in these experiments involved different rates for distinct parts of the network. For the dual and triple encoder frameworks, pre-trained models used a learning rate of $1e-6$, enabling gradual feature learning, as observed in the baseline models. In contrast, the MLP encoders performed better with a learning rate of $1e-5$, based on findings from Subsection 3.3.2 and Subsection 3.3.3. Consequently, the final setup applied a learning rate of $1e-6$ to the pre-trained encoders and $1e-5$ to other learnable components.

3.4 Evaluation Metrics

The downstream task in this work was multi-label classification, which required performance metrics to evaluate the model's predictions at two levels. Balanced Accuracy was used to evaluate the model's ability to correctly predict the entire multi-label set, providing a stricter measure of performance. On the other hand, the F1 Score was used to assess the model's ability to correctly predict any subset of the multi-label set, offering a more lenient evaluation.

3.4.1 Balanced Accuracy (BA)

Balanced Accuracy (BA) is computed by macro-averaging the accuracy over the unique multi-label patterns. Rather than treating each pattern as a separate class, the metric calculates the accuracy for each distinct label set and then averages these values. This approach ensures that every unique label combination contributes equally to the overall performance metric, thereby mitigating the effects of class imbalance. Mathematically, it is defined as:

$$\text{Pattern-wise Balanced Accuracy} = \frac{1}{|P|} \sum_{p \in P} \frac{C_p}{N_p}$$

where $|P|$ is the number of unique label patterns, C_p represents the correct predictions for pattern p, and N_p represents the total instances of pattern p. A good BA score will imply that the model is good at predicting the entire multi-label pattern.

3.4.2 F1 Score

The F1 score or label-based F1 score [54] evaluates model performance by calculating the F1 score independently for each class and averaging these scores. This metric also ensures the incorporation of the class imbalance while scoring the predictions. A high F1 score reflects the model's ability to predict individual classes accurately. The mathematical representation for this metric is as follows:

$$F1 = \frac{1}{C} \sum_{i=1}^{C} \frac{2 \cdot \text{TP}_i}{2 \cdot \text{TP}_i + \text{FP}_i + \text{FN}_i}$$

where C represents the total number of classes in the dataset. TP_i denotes the number of true positives for class i, FP_i refers to the number of false positives for class i and, FN_i represents the number of false negatives for class i.

3.4.3 Dominance Score

To provide a robust method for comparing models across multiple metrics, we propose a novel dominance score. Although this metric is not a central focus of our work, it offers a convenient way to quantify how consistently a model outperforms its competitors when various evaluation criteria are considered.

The dominance score evaluates the overall performance of models across CATAMI hierarchical levels and sonar data types (backscatter and bathymetry). For each depth, a model's performance is assessed at specific confidence levels (mean, one standard deviation, and two standard deviations) for metrics (Balanced Accuracy and F1). A model earns one point if its performance exceeds all others for a given depth, confidence level, and metric. This is repeated across all levels, sonar types, metrics, and confidence levels. The total dominance score reflects the cumulative points awarded to a model. This metric was used to identify the most effective fusion framework, addressing the second sub-research question, and was applied in experiments 2 and 3.

The mathematical formulation, where DS_i reflects the total number of configurations where model i outperforms all others, is:

$$DS_i = \sum_{d \in D} \sum_{s \in S} \sum_{m \in M} \sum_{c \in C} \mathbb{I}\big(P_i(m,c,d,s) > P_j(m,c,d,s)\ \forall\, j \neq i\big)$$

where $D = \{d_1, d_2, d_3\}$ represents CATAMI hierarchical levels, $S = \{s_1, s_2\}$ sonar data types[c], $M = \{m_1, m_2\}$ evaluation metrics, and $C = \{c_1, c_2, c_3\}$ confidence levels. $P_i(m,c,d,s)$ denotes the performance of model i for a specific combination of m, c, d, s.

The indicator function $\mathbb{I}(x)$ is defined as:

$$\mathbb{I}(x) = \begin{cases} 1 & \text{if } x \text{ is true,} \\ 0 & \text{otherwise.} \end{cases}$$

[c]This variable was only used in Experiment 2; in Experiment 3, where both the sonar data types were used together, the dominance score was computed without this variable.

Chapter 4

Results and Discussion

4.1 Spatial Feature Experiments

The results of the baseline BayesianKDE classifier and Random forest models, using single-point and context-based sampling, which compares the F1 and BA scores across the three CATAMI hierarchy levels, are illustrated in Figure 4.1. From these results, three distinct observations emerge.

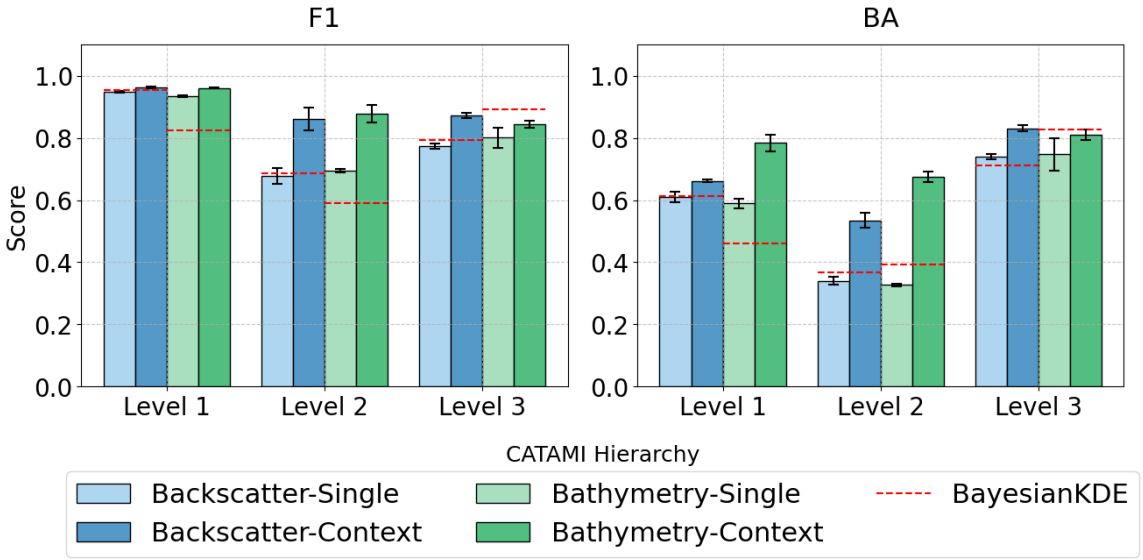

Figure 4.1: The bar plots illustrate the F1 scores (left) and BA scores (right) of Random Forest models across all the CATAMI hierarchical level (Values presented in numerical form in Table A.1). The light blue colour is used to represent the model trained on single-point data, and the darker blue colour is given for the model trained on context-based data. Similarly, the light green and the dark green bars represent the RF models trained on single-point and context-based bathymetry data, respectively. The red dashed line illustrates the performance set by the BayesianKDE model at the given CATMI depth (Values presented in numerical form in Table A.8). depict the mean score of the Random forest model after the 5-fold cross-validation, and the error bars show the variance.

First, at CATAMI hierarchical levels 1 and 2, the baseline BayesianKDE classifier shows strong performance when using single-point backscatter data, reflecting the

high correlation between individual backscatter values and substrate classes. However, when applied to bathymetry data, BayesianKDE's performance declines noticeably, which suggests that bathymetry has a more complex and less direct relationship with substrate type. In contrast, the Random Forest model leverages its ensemble of decision trees to partition the feature space, enabling it to capture these complex relationships more effectively. This allows the RF model trained on single-point bathymetry data to outperform the BayesianKDE baseline at Levels 1 and 2, which is not the case with backscatter data. Notably, at CATAMI hierarchical level 3, where two labels consistently co-occur (explained in Subsection 2.2.2), the BayesianKDE classifier outperforms the Random Forest model when using bathymetry data. In our dataset, these labels are encoded as [1,1,0] (for the multilabel class of Coarse sand-gravel) and [0,0,1] (for the class Mud/slit). However, the Random Forest model may predict an output such as [1,0,0], which does not exist in the dataset, thereby introducing additional misclassifications by not fully capturing the inherent interdependence between these labels. In contrast, the BayesianKDE approach treats the labels at this depth as binary classes, effectively bypassing this issue. It is important to note, however, that if the labels at Level 3 did not co-occur so consistently, the performance advantage observed for BayesianKDE might diminish.

Second, at CATAMI hierarchical level 2, the overall performance of all models declines, a trend that aligns with the increasing complexity of the label space as more substrate classes are introduced (illustrated in Figure 2.3). The more challenging distinctions at this depth naturally lead to reduced classification accuracy.

Finally, a unifying observation is that context-based sampling consistently enhances classification performance across levels. Incorporating adjacent sonar pixels enriches the local spatial signature and provides complementary information that refines the visual features used for classification. This improvement is particularly notable in scenarios with more complex label distributions. In essence, the enhanced performance achieved with context-based sampling directly answers our research question by demonstrating that wider spatial context significantly benefits substrate classification in the Bay of Fundy region.

4.2 Dual and Triple Encoder Fusion Experiments

4.2.1 Dominance Score Analysis for Dual-Encoder Architecture

The dominance scores of the SMR (Shallow marginal representations) and DMR (Deep marginal representations) models in a dual-encoder setting are given in Table 4.1. The SMR model consistently achieves higher dominance scores across all confidence levels, outperforming the DMR model. Specifically, SMR exhibits superior generalization and robustness, achieving a total of 8 dominance scores at the 2-σ level, while the DMR model records no wins.

Model	μ	**1-σ**	**2-σ**
SMR	**10**	**9**	**8**
DMR	2	0	0

Table 4.1: Dominance scores of the dual-encoder models across different confidence levels. In the case of this architecture, the scores are given out of 12. The numerical values of the performance metrics are given in Table A.2 and Table A.3. The 2-σ dominances are highlighted in bold in these tables.

These findings indicate that, for dual-encoder fusion, the feature representation required for sonar data (backscatter and bathymetry) is relatively straightforward. A single-layer MLP (SMR) is sufficient to capture the essential characteristics of both backscatter and bathymetry, while deeper representations appear to introduce unnecessary complexity and risk over-parameterization.

4.2.2 Dominance Score Analysis for Triple-Encoder Architecture

In contrast to the dual-encoder setup, the triple-encoder models show similar performance. However, the bsSMR-bmDMR configuration, which combines a shallow representation for backscatter with a deeper representation for bathymetry, emerges as slightly more favourable due to it achieving a dominance score of 1 at the 2-*sigma* confidence level.

Model	μ	1-σ	2-σ
bsSMR-bmSMR	1	0	0
bsDMR-bmDMR	**3**	1	0
bsSMR-bmDMR	2	1	**1**
bsDMR-bmSMR	0	0	0

Table 4.2: Dominance scores of the dual-encoder models across different confidence levels. In the case of this architecture, the scores are given out of 8. The numerical values of the performance metrics are given in Table A.4.

The slight improvement observed with the bsSMR-bmDMR model is most evident at CATAMI hierarchical level 2[a], which has a more complex label space. This result suggests that when substrate classes become more challenging to differentiate, a deeper abstraction of bathymetry can provide a visible performance improvement. Additionally, the findings from Hasan et al.[55] further imply that careful feature selection or abstraction is beneficial for bathymetry, supporting our choice of using DMR for that sonar data type.

4.2.3 Performance Metrics Evaluation

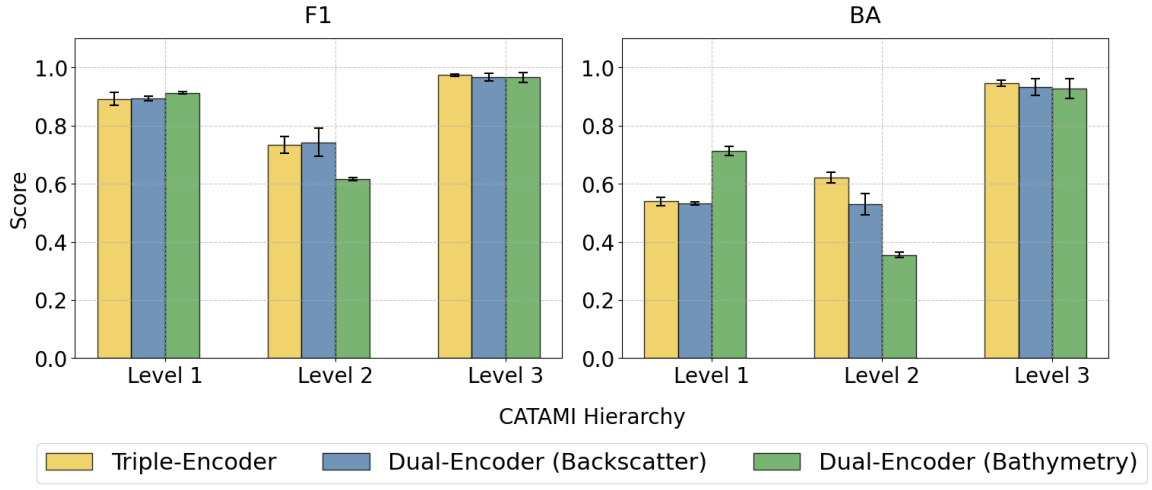

Figure 4.2: Comparison of the F1 (left) and BA scores (right) of the bsSMR-bmDMR model (yellow bar) with the SMR dual-encoder model trained on backscatter (blue bar) and bathymetry (green bar) across all the CATAMI hierarchy.

The comparison of performance metrics for the dual-encoder SMR models and the

[a]Observation is highlighted in Table A.4.

triple-encoder bsSMR-bmDMR is shown in Figure 4.2. The backscatter-based and bathymetry-based SMR models perform similarly overall, though subtle differences emerge across the CATAMI hierarchy. At Level 1, the dual-encoder bathymetry model shows a notable boost in Balanced Accuracy (BA) compared to the other models, suggesting that these substrate classes become more distinguishable when bathymetry is fused with image features. Conversely, both the dual-encoder models experience a dip in BA at Level 2, reflecting the complexity of the label space at this level.

Comparing the dual-encoder models with the triple-encoder bsSMR-bmDMR framework reveals generally comparable BA and F1 scores at most levels. However, at Level 2, the bsSMR-bmDMR model is over 2-σ significant in BA scores compared to the dual-encoder models, consistent with the notion discussed earlier that more complex label spaces benefit from deeper or more nuanced feature representations. Interestingly, the backscatter-based dual-encoder and triple-encoder models yield statistically similar F1 scores across all the CATAMI hierarchical levels, indicating that simpler architectures can still perform competitively when predicting independent classes. These findings underscore that while the triple-encoder architecture can provide richer representations in complex scenarios, dual-encoder models remain strong contenders in simpler contexts with fewer parameters.

4.2.4 Summary of Findings

Overall, our experiments reveal that improvements observed in dual-encoder setups do not extend to triple-encoder setups. This behaviour can be attributed to the triple-encoder setup working with both the data types of the sonar modality compared to the dual-encoder models that work with only one, thus requiring a deeper representation of the bathymetry data to capture the intermodal relationships. Despite that, these experiments address the second sub-research question by identifying effective fusion strategies for multimodal seafloor substrate classification. For dual-encoder models, the SMR framework effectively integrates sonar data with visual features. For triple-encoder models, although performance differences among configurations are minimal, we selected the bsSMR-bmDMR configuration for its consistent performance in more challenging scenarios, particularly at CATAMI hierarchical level 2.

4.3 Pre-trained encoder-based fusion experiments

4.3.1 Performance Metrics Evaluation

The performance of the baseline pre-trained ResNet50 and Vision Transformer (ViT) models, each fine-tuned on benthic image data alone, is shown in Figure 4.3. Both models achieve high F1 scores, demonstrating their capacity to correctly classify individual classes. However, ResNet50 models achieve higher BA scores compared to ViT, specifically at CATAMI hierarchical level 1 and depth 3. In contrast, the relative underperformance of the ViT model may be attributed to its reliance on larger datasets during fine-tuning and the absence of intrinsic spatial inductive biases that are typical of convolutional networks [27].

A comparison of the ResNet50 and Vit-based dual and triple encoder models is also shown in Figure 4.3. For ResNet50-based models, neither dual-encoder nor triple-encoder configurations show statistically significant improvements over the baseline at any CATAMI hierarchical level. By contrast, the ViT-based fusion models exhibit pronounced gains across all levels, with BA scores improving significantly over the ViT baseline—particularly in the triple-encoder framework at CATAMI hierarchical level 2. Direct comparisons between ViT and ResNet50-based fusion models should be made cautiously due to differences in parameter counts and architectural biases.

Xu et al. [56] emphasize that the transformer-based model's tokenization process treats each patch as an independent token. This approach aligns with the concept of feature-level fusion, as the ViT outputs a sequence of embeddings that can directly interact with other modalities. In contrast, ResNet models produce feature maps that may not work well for cross-modal fusion. This paper provides theoretical grounding for the behaviour of the ViT model and highlights the potential of transformer-based models in multimodal settings. An ablation study was done on these multimodal models to gain more insights.

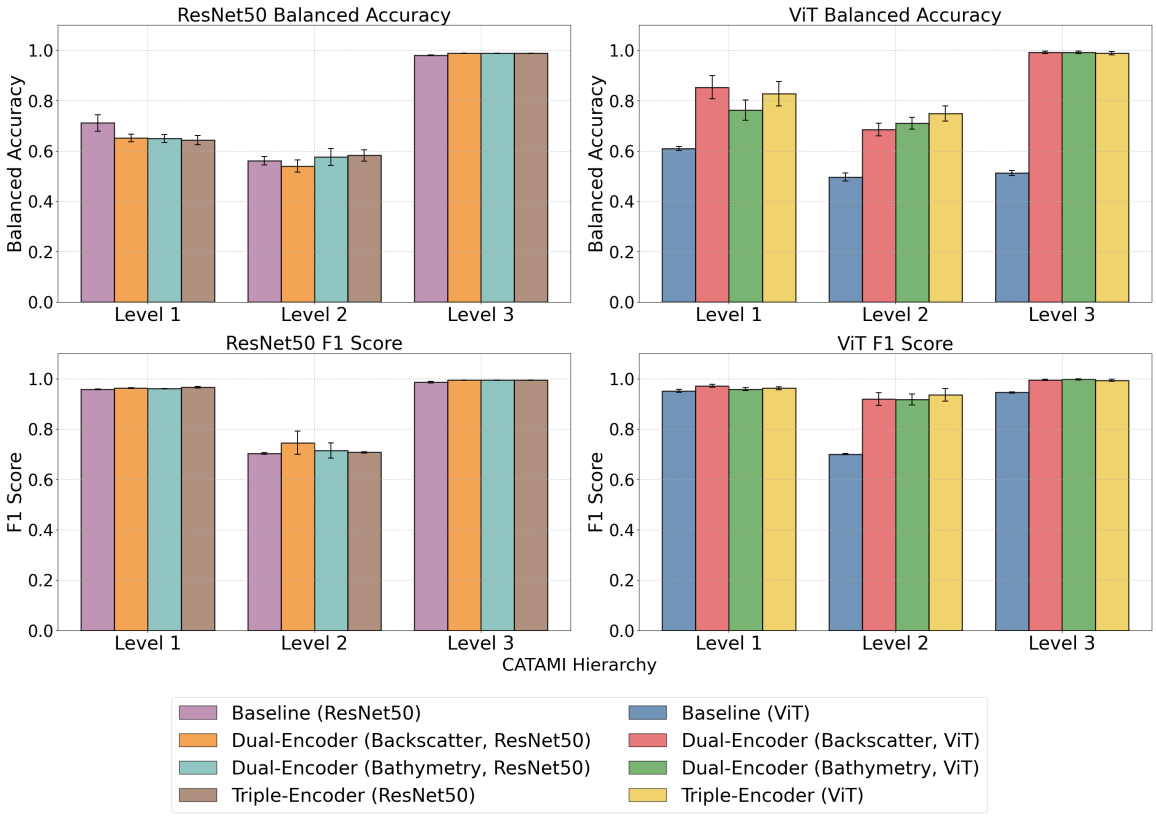

Figure 4.3: F1 score and BA score of the ResNet50 (left) and ViT (right) for baseline and multimodal models are illustrated in this graph. The baseline ResNet50 models are represented by purple bars, while the baseline ViT models are depicted with blue bars. The orange and teal bars represent the ResNet50-based dual-encoder models trained using backscatter and bathymetry, respectively. Similarly, the red and green colours denote the ResNet50-based dual-encoder models trained with backscatter and bathymetry as well. The triple encoder ResNet50 model is shown with brown bars, and the triple encoder model of ViT appears in the graph yellow.

4.3.2 Ablation Study

In this ablation study, we aimed to evaluate the contribution of each modality within our fusion framework. To conduct this study, we performed linear probes on the individual encoders extracted from dual and triple-encoder models trained in a multimodal setting. This investigation examined how the fusion of sonar and image data influences classification performance and determined whether the sonar encoders contribute independently to the task or primarily serve to enhance the visual representations extracted by the image encoders.

The results reveal important insights[b] into the role of each encoder when evaluated

[b]Tables are given in Table A.9 and Table A.10

independently. Notably, the performance metrics for the individual sonar encoders (backscatter and bathymetry) do not show significant improvements when incorporated into the fusion model.

Performance change in the image encoders of the ResNet50 and ViT-based dual and triple-encoder models, compared to the baseline ResNet50 and ViT models, is illustrated in Figure 4.4. The ViT-based image encoder's performance (shown in graphs 2, 4, 6), measured by Balanced Accuracy (BA) and F1 metrics, when fused with sonar data in both dual and triple-encoder models, improves significantly when compared to baseline ViT. Specifically, the BA scores show an improvement of at least 1-σ across all the CATAMI hierarchical levels. This enhancement indicates that the integration of sonar data allows the ViT encoder to extract more discriminative features than ResNet50. This pattern is consistent across all the ViT-based multimodal models, underscoring the complementary nature of the sonar data in augmenting visual representations and resolving ambiguity in visually similar substrates.

In contrast, the ResNet50-based models exhibit much smaller gains from including sonar data, as shown in graphs 1, 3 and 5. In some cases, the BA and F1 scores remain near baseline levels or show slight declines, particularly at CATAMI hierarchical levels 1 and 2. This performance suggests that the ResNet50 encoder does not exploit the complementary information provided by sonar data as effectively as the ViT encoder. The triple-encoder models reinforce this similar trend between ResNet50 and ViT.

Summary of Findings: Ablation Study

The results indicate that while the standalone performance of the sonar encoders is modest, their integration substantially enhances the image encoder's discriminative power. The ViT image encoder benefits markedly from including sonar data, with all the multimodal image encoders having significant improvements compared to their baselines, while the ResNet50 doesn't. The results of the ablation tests suggest that the ViT-based multimodal models get significant gains, leading to a fine-grained feature representation for the classes.

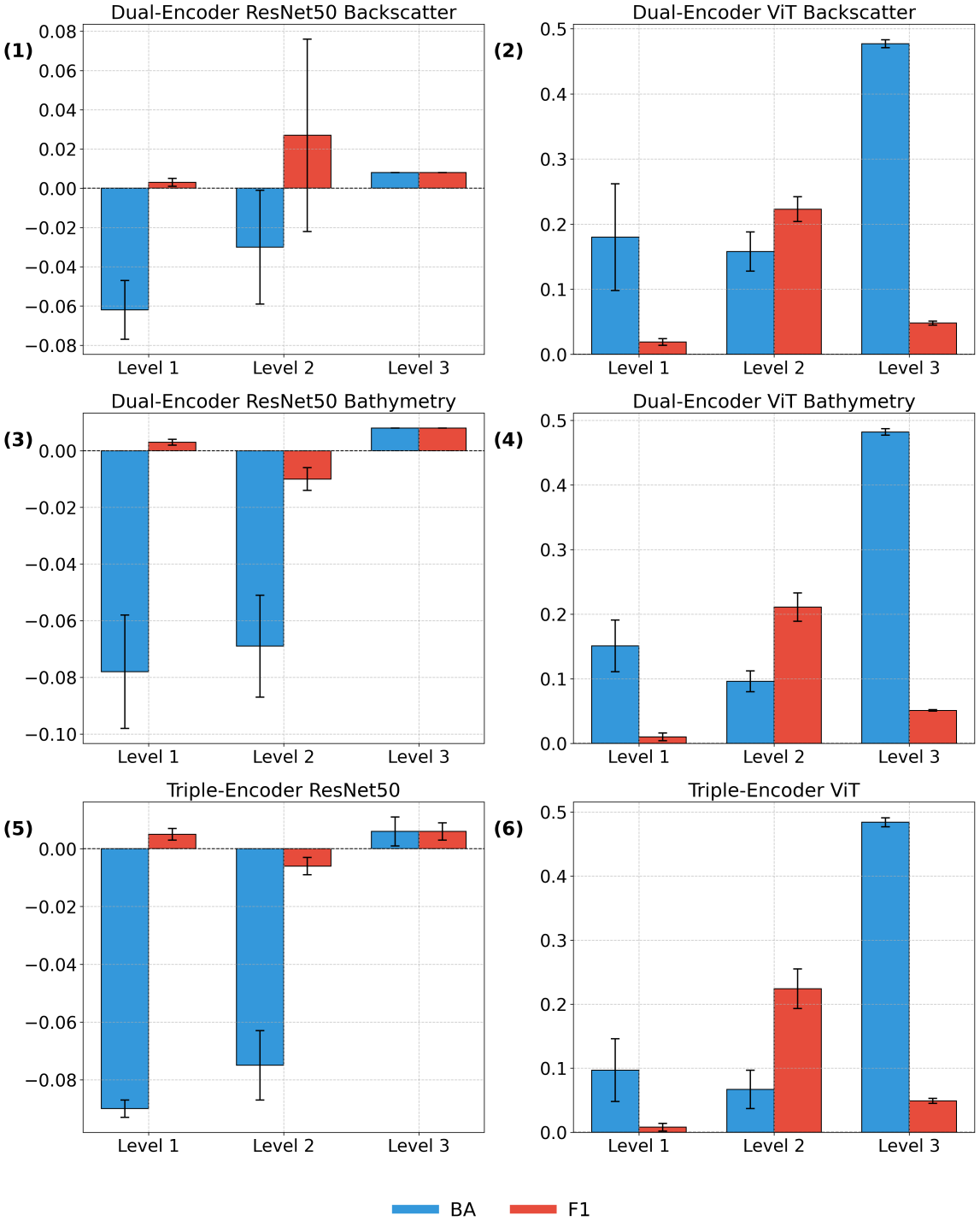

Figure 4.4: The F1 and BA scores of individual image encoders from the ablation study, along with baseline models (ResNet50 and ViT) fine-tuned solely on image data, are presented in the graphs. F1 scores are shown in red bars, while BA scores are in blue bars. Graphs 1 and 3 display the changes in performance for ResNet50-based dual-encoder models trained on backscatter and bathymetry data, respectively. In contrast, graphs 2 and 4 depict the same changes for ViT-based models. Graph 5 illustrates the performance changes for ResNet50-based triple encoder models, while Graph 6 shows the same for ViT-based models.

4.3.3 Feature Space Visualization

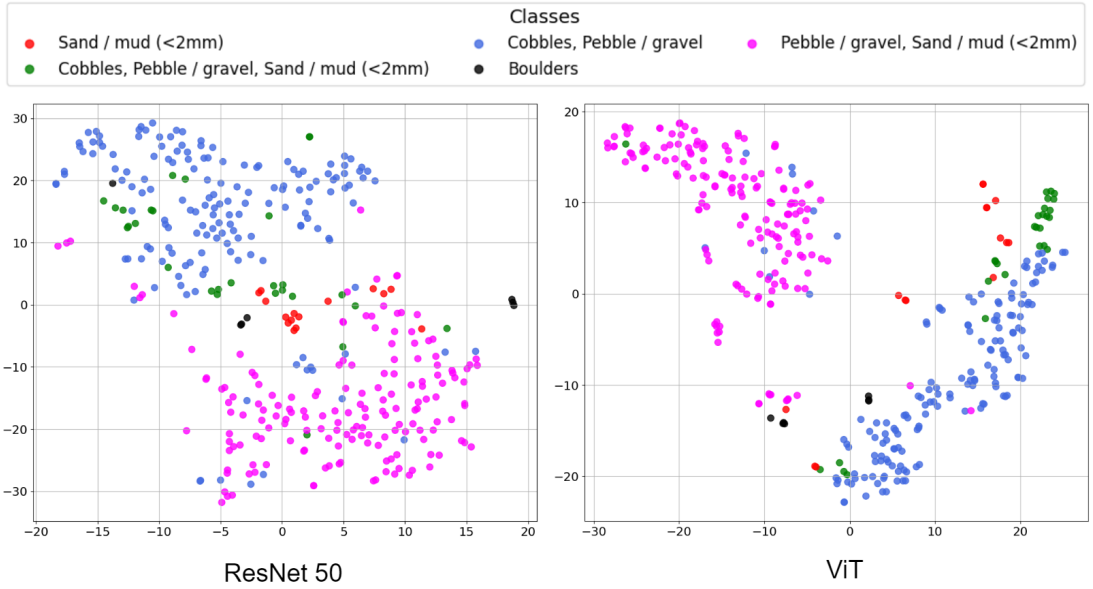

Figure 4.5: t-SNE plots of the triple-encoder ResNet50 (left) and ViT (right) models at CATAMI hierarchical level 2. The majority classes, Pebble/Grave-Sand/Mud (<2mm) and Cobbles-Pebble/gravel, are represented in pink and blue, respectively. The minority class, Sand/Mud (<2mm), is shown in red. Another minority class, Cobbles-Pebble/Gravel-Sand/Mud (<2mm), is depicted in green, while the minority class of Boulders is represented in black.

The t-SNE visualizations of the learned feature space are illustrated in Figure 4.5 [c]. Although these low-dimensional projections do not fully represent the high-dimensional feature space, they illustrate that ViT-based models exhibit more distinct clustering of substrate classes. In particular, the ViT representations show clearer separation between both majority and minority classes, such as the Sand/Mud (<2mm) and Cobbles/Pebble-Gravel clusters, compared to ResNet50-based models. This observation further reinforces the advantage of transformer-based architectures in leveraging complementary sonar information.

[c]The plots of other models at different CATAMI hierarchical levels showed similar plot characteristics.

4.3.4 Comparing Triple-encoder ViT with Context-based Random Forest Model

The triple encode ViT model is compared with the random forest model trained on context-based features of sonar data in Figure 4.6. This comparison was done to evaluate the performance difference between a simpler random forest model trained on sonar data and a multimodal model. Here, we observe that the triple-encoder ViT model, which generally performed better than other multimodal models, outperforms the RF models at CATAMI hierarchical level 2 and 3 by over two standard deviations in BA scores, and it achieves approximately one standard deviation better score in other levels.

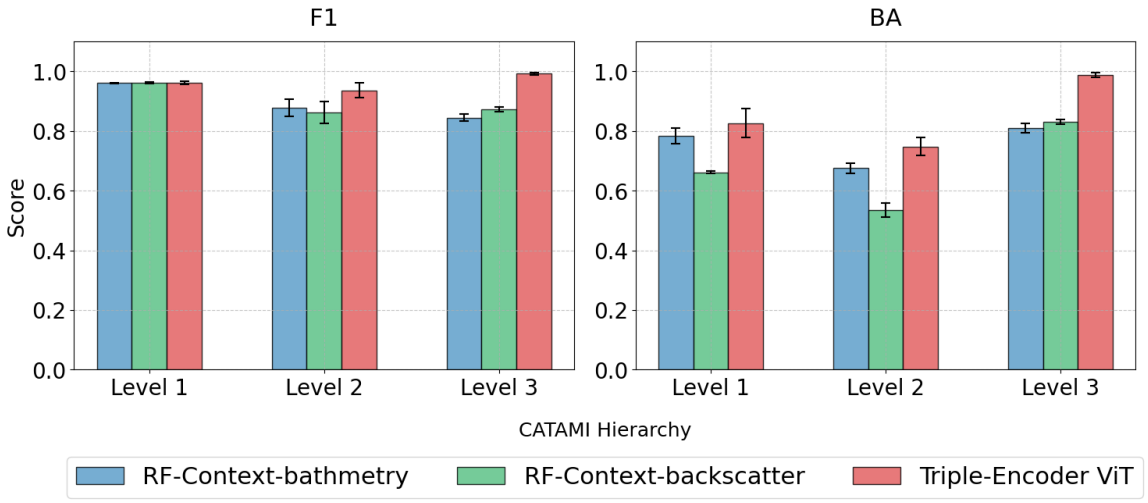

Figure 4.6: Comparison of F1 (left) and BA scores (right) of the ViT-Based Triple-encoder model (red bars) with the Context-based Random first model trained on backscatter (blue) and bathymetry (green).

Regarding RF models, the F1 scores are comparable to triple-encoder ViT models at levels 1 and 2, indicating that simpler models are competitive with the multimodal model to predict the independent classes in the multi-label set. This may be because the random forest model is able to identify decision boundaries that effectively separate individual classes, similar to the multimodal model. However, these boundaries do not fully capture the nuances of the complete multilabel sets, leading to a performance deficit in BA scores of approximately 2-σ compared to the multimodal ViT model across CATAMI hierarchical levels, with the exception of Level 1. These results suggest that when the use case is to predict these independent classes, choosing the

random forest model will be beneficial due to its lighter computational requirements. While multimodal models will be useful for multi-label prediction tasks.

4.3.5 Summary of Findings

Overall, our findings answer the third sub-research question here. When generalizing the pre-trained models on the selected fusion architecture from Experiment 2 (Subsection 3.3.2) and 3 (Subsection 3.3.3), ViT-based multimodal models show substantial performance improvements, while ResNet50-based fusion yields only modest gains. These results support the usefulness of the fusion process in enhancing image encoder performance and suggest that transformer-based models may hold promise for multimodal benthic substrate classification.

Chapter 5

Conclusion, Limitations and Future Works

5.1 Conclusion

In this thesis, research was conducted to determine whether multimodal learning, through the addition of sonar data, improves substrate image classification in the Bay of Fundy region. Three sub-research questions were addressed to explore this overarching goal:

1. Does using a wider spatial context for the sonar data help with the benthic substrate classification? To investigate this question, we experimented with two data sampling methods, namely, single-point (50 x 50 m^2 spatial coverage) and context-based sampling (150 x 150 m^2 spatial coverage). Random Forest models trained with context-based features significantly outperformed both the baseline BayesianKDE model and models trained with single-point features, highlighting the importance of wider spatial context in benthic substrate classification. These findings suggest that context-based sampling is a more effective method for extracting and utilizing sonar data features, making it a valuable approach for improving the accuracy of seafloor classification tasks.

2. Which feature-level fusion framework is the most effective for multimodal seafloor substrate classification? In this study, we evaluated both the dual-encoder and triple-encoder multimodal architectures by testing two approaches for feature extraction from sonar data, which are shallow (SMR) and deep (DMR). For the dual-encoder framework, our experiments demonstrated that models using a shallow marginal representation (SMR) for sonar data outperformed those employing deeper representations, indicating that the intrinsic complexity of individual sonar modalities is low enough to be effectively captured with minimal transformation. Conversely, a more nuanced picture emerged in the triple-encoder framework, where visual data is fused with both backscatter and bathymetry. Although overall performance across triple-encoder variations was comparable, the bsSMR-bmDMR model, which uses

a shallow representation for backscatter and a deep representation for bathymetry, exhibited a little statistical significance with improvement in Balanced Accuracy at CATAMI hierarchical level 2, where the label space is particularly complex. Supported by theoretical insights, we selected the bsSMR-bmDMR model framework for further experimentation with triple-encoder architecture.

3. Can the proposed multimodal framework generalize effectively to pre-trained image networks? The dual and triple-encoder architectures, which were found to be effective based on the findings of the second sub-research question, were employed with the image encoder replaced by pre-trained ResNet50 or Vision Transformer (ViT) models. The experiments demonstrate that the multimodal framework generalizes effectively to pre-trained image networks, though its benefits vary by architecture. Specifically, ViT-based fusion models consistently leverage the addition of sonar data—evidenced by significant gains across all the CATAMI hierarchical levels compared to its baseline (ViT fine-tuned on image modality) and show tighter class clustering in t-SNE visualizations. In contrast, ResNet50-based fusion yields only modest improvements under limited conditions, suggesting that its convolutional design may be less suited for cross-modal feature integration. Overall, these findings highlight the promise of transformer-based architectures for enhancing image encoder performance through complementary sonar information while also indicating that further investigation is needed to fully understand and optimize the fusion mechanisms at play.

In conclusion, this research has contributed to the field of marine substrate classification by demonstrating the effectiveness of multimodal learning approaches. By investigating data sampling methods and fusion techniques, we established a robust framework for combining visual and sonar data. The enhanced performance of multimodal models confirms that the addition of sonar data enhances the classification capabilities of models for seafloor substrate images in the Bay of Fundy dataset. These findings not only advance the technical aspects of seafloor classification but also have practical implications for marine habitat mapping and conservation efforts, highlighting the potential of multimodal frameworks for improving marine resource management.

5.2 Limitations and Future Directions

While this thesis demonstrates the promise of multimodal fusion for benthic substrate classification, several constraints temper the generality of our findings and point toward opportunities for advancement.

5.2.1 Dataset Constraints

A key limitation lies in the geographic specificity of our dataset, which is drawn exclusively from the Bay of Fundy. Its distinctive sediment composition, tidal dynamics, and clear backscatter and bathymetry contrasts may not be representative of other marine environments. Consequently, the fusion strategies validated here may not translate easily to regions where acoustic–substrate relationships are more subtle or where sediment distributions are more heterogeneous.

In addition to this, the spatial scales of our two modalities differ markedly: each sonar pixel covers a 50 m $\times$ 50 m seafloor area, whereas images capture merely a 1–2 m footprint. Even with context-based sampling, this scale mismatch leads multiple images to inherit identical sonar values, risking both overfitting and the loss of fine-scale substrate heterogeneity. Moreover, each underwater image can carry a different composition of multiple labels (for example, [90% gravel, 10% sand] versus [10% gravel, 90% sand]) yet be treated as a single multi-label class. This simplification obscures compositional nuances that critically shape habitat interpretation and backscatter signatures.

Although some images include scaling lasers in their field of view, the majority of images in the dataset do not. In our work, we did not filter out images containing these lasers; thus, this information was not well-represented in the training data. Scaling lasers provide a direct reference for object and sediment size in each photograph, aiding in distinguishing fine sediments such as mud or silt. Since most images did not have scaling lasers, the models may not have learned the size-dependent features crucial for classifying certain substrate classes, limiting their performance.

In addition, we did not derive slope information from the bathymetry data. Seafloor slope can be a key predictor of substrate class, as sharp gradients often coincide with changes in sediment grain size or habitat type. Omitting slope thus

constrains the richness of our acoustic feature set.

Finally, our dataset lacks temporal and environmental metadata such as seasonal variation, temperature and salinity. These factors can influence sediment properties and acoustic responses over time, and their absence prevents models from capturing dynamic habitat changes.

5.2.2 Methodological Constraints

Our investigation is further limited by its exclusive focus on feature-level fusion. Early fusion of raw inputs or late fusion of model decisions may uncover complementary patterns or better mitigate modality-specific noise, but these approaches remain unexplored here. Additionally, we adopt a fixed 3×3 sonar-pixel window (150 m square) for context sampling, which may prove suboptimal in zones of abrupt substrate change. An adaptive or multi-scale sampling strategy could more faithfully capture local heterogeneity and reduce label noise. Although our image encoders leveraged large-scale pre-training (ResNet50, ViT), sonar encoders (MLPs) were trained from scratch, limiting their ability to extract rich acoustic features. Moreover, we evaluated only a limited set of visual backbones, such as CNN, ResNet50, and ViT, leaving newer architectures and multimodal transformer networks unexplored. Finally, we did not incorporate temporal or environmental metadata (for example, seasonal shifts, temperature, or salinity), nor did we employ hierarchical multi-label classifiers that model label co-occurrence or dependency, each of which could yield deeper insights.

5.2.3 Future Directions

Future studies should extend to datasets from diverse marine regions and seabed types to address geographic bias. Adaptive or multi-scale sonar sampling, where the context window size varies with local substrate heterogeneity, could reduce label noise and better capture relevant acoustic signatures. Pre-training sonar encoders on large acoustic corpora, perhaps via self-supervised learning, may yield stronger marginal representations.

Additionally, consistently filtering images to ensure scaling lasers are represented might enable models to accurately learn true object and sediment dimensions, thereby improving the classification of fine sediments. Experiments can also be conducted

with a set of images that contain no laser information at all to evaluate the difference this metadata makes. Deriving slope and other terrain metrics from bathymetry data can provide valuable predictors of substrate transitions. Moreover, pre-training sonar encoders on large acoustic datasets, possibly through self-supervised learning, could yield stronger marginal representations.

Furthermore, exploring early fusion of raw sonar and image data, as well as late fusion ensembles, could reveal complementary information overlooked by feature-level fusion. Testing advanced visual backbones (e.g., Swin Transformer, ConvNeXt) and emerging multimodal transformer frameworks may further enhance cross-modal integration. Additionally, incorporating temporal metadata (season, tide) and environmental variables (temperature, salinity) could capture dynamic habitat changes, while hierarchical multi-label models would more fully exploit CATAMI's taxonomy. Addressing these avenues promises to strengthen generalizability, accuracy, and practical impact in multimodal oceanographic mapping.

Bibliography

[1] H. Mohamed, K. Nadaoka, and T. Nakamura, "Assessment of machine learning algorithms for automatic benthic cover monitoring and mapping using towed underwater video camera and high-resolution satellite images," *Remote Sensing 2018, Vol. 10, Page 773*, vol. 10, p. 773, 5 2018. [Online]. Available: https://www.mdpi.com/2072-4292/10/5/773/

[2] A. Gómez-Ríos, S. Tabik, J. Luengo, A. S. Shihavuddin, B. Krawczyk, and F. Herrera, "Towards highly accurate coral texture images classification using deep convolutional neural networks and data augmentation," *Expert Systems with Applications*, vol. 118, pp. 315–328, 3 2018. [Online]. Available: https://arxiv.org/abs/1804.00516v1

[3] G. Li, F. Wang, L. Zhou, S. Jin, X. Xie, C. Ding, X. Pan, and W. Zhang, "Mcanet: Multi-channel attention network with multi-color space encoder for underwater image classification," *Computers and Electrical Engineering*, vol. 108, p. 108724, 5 2023. [Online]. Available: https://doi.org/10.1016/j.compeleceng.2023.108724

[4] A. Noor and N. I. Ruhaiyem, "Underwater image processing based on cnn applications: A review," *Proceedings of the Cognitive Models and Artificial Intelligence Conference*, pp. 75–84, 5 2024. [Online]. Available: https://dl.acm.org/doi/10.1145/3660853.3660870

[5] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *Transactions on Pattern Analysis and Machine Intelligence*, 2018. [Online]. Available: http://www.ieee.org/publications_standards/publications/rights/index.html

[6] S. C. Lowe, B. Misiuk, I. Xu, S. Abdulazizov, A. R. Baroi, A. C. Bastos, M. Best, V. Ferrini, A. Friedman, D. Hart, O. Hoegh-Guldberg, D. Ierodiaconou, J. Mackin-Mclaughlin, K. Markey, P. S. Menandro, J. Monk, S. Nemani, J. O'brien, E. Oh, L. Y. Reshitnyk, K. Robert, C. M. Roelfsema, J. A. Sameoto, A. C. G. Schimel, J. A. Thomson, B. R. Wilson, M. C. Wong, C. J. Brown, and T. Trappenberg, "Benthicnet: A global compilation of seafloor images for deep learning applications," *Scientific Data*, 2024. [Online]. Available: https://doi.org/10.20383/103.0614.

[7] K. Bayoudh, R. Knani, F. Hamdaoui, A. Mtibaa, B. K. Bayoudh, F. Hamdaoui, and A. Mtibaa, "A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets," *The Visual Computer 2021 38:8*, vol. 38, pp. 2939–2970, 6 2021. [Online]. Available: https://link.springer.com/article/10.1007/s00371-021-02166-7

[8] J. Li, D. Hong, L. Gao, J. Yao, K. Zheng, B. Zhang, and J. Chanussot, "Deep learning in multimodal remote sensing data fusion: A comprehensive review," *International Journal of Applied Earth Observation and Geoinformation*, vol. 112, p. 102926, 8 2022. [Online]. Available: https://doi.org/10.1016/j.jag.2022.102926

[9] G. Muhammad, F. Alshehri, F. Karray, A. E. Saddik, M. Alsulaiman, and T. H. Falk, "A comprehensive survey on multimodal medical signals fusion for smart healthcare systems," *Information Fusion*, vol. 76, pp. 355–375, 12 2021. [Online]. Available: https://doi.org/10.1016/j.inffus.2021.06.007

[10] M. Wang, D. Wang, Y. Xiang, Y. Liang, R. Xia, J. Yang, F. Xu, and X. Huang, "Fusion of ocean data from multiple sources using deep learning: Utilizing sea temperature as an example," *Frontiers in Marine Science*, vol. 10, p. 1112065, 2 2023. [Online]. Available: https://doi.org/10.3389/fmars.2023.1112065

[11] D. Rao, M. D. Deuge, N. Nourani-Vatani, B. Douillard, S. B. Williams, and O. Pizarro, "Multimodal learning for autonomous underwater vehicles from visual and bathymetric data," *Proceedings - IEEE International Conference on Robotics and Automation*, pp. 3819–3825, 9 2014. [Online]. Available: https://doi.org/10.1109/ICRA.2014.6907413

[12] N. Oceanic and A. A. U. D. of Commerce, "What is sonar?" [Online]. Available: https://oceanservice.noaa.gov/facts/sonar.html

[13] D. Neupane and J. Seok, "A review on deep learning-based approaches for automatic sonar target recognition," *Electronics 2020, Vol. 9, Page 1972*, vol. 9, p. 1972, 11 2020. [Online]. Available: https://www.mdpi.com/2079-9292/9/11/1972

[14] H. Bahrami and S. R. Talebiyan, "Sonar signal classification using neural networks," *Canadian Conference on Electrical and Computer Engineering*, 2015. [Online]. Available: https://doi.org/10.1109/CCECE.2000.849558

[15] "Exploration tools: Multibeam sonar: Noaa office of ocean exploration and research." [Online]. Available: https://oceanexplorer.noaa.gov/technology/sonar/multibeam.html

[16] V. Renard and J.-P. Allenou, "Sea beam, multi-beam echo-sounding in "jean charcot" - description, evaluation and first results," *The International Hydrographic Review*, 1979. [Online]. Available: https://journals.lib.unb.ca/index.php/ihr/article/view/23653

[17] C. J. Brown and P. Blondel, "Developments in the application of multibeam sonar backscatter for seafloor habitat mapping," *Applied Acoustics*, vol. 70, pp. 1242–1247, 10 2009. [Online]. Available: https://doi.org/10.1016/J.APACOUST.2008.08.004

[18] B. J. Todd, J. M. Shaw, D. R. Parrott, J. E. H. Clarke, D. Cartwright, and S. E. Hayward, "Backscatter strength and shaded seafloor relief, bay of fundy, sheet 1, offshore nova scotia - new brunswick," 11 2011. [Online]. Available: https://ostrnrcan-dostrncan.canada.ca/handle/1845/162012

[19] B. J. Todd, J. M. Shaw, and D. R. Parrott, "Shaded seafloor relief, bay of fundy, sheet 1, offshore nova scotia - new brunswick," 11 2011. [Online]. Available: https://ostrnrcan-dostrncan.canada.ca/handle/1845/161276

[20] B. Misiuk, C. J. Brown, K. Robert, and M. Lacharité, "Harmonizing multi-source sonar backscatter datasets for seabed mapping using bulk shift approaches," *Remote Sensing 2020, Vol. 12, Page 601*, vol. 12, p. 601, 2 2020. [Online]. Available: https://www.mdpi.com/2072-4292/12/4/601

[21] C. Haar, B. Misiuk, V. Gazzola, M. Wells, and C. J. Brown, "Harmonizing multi-source backscatter data using bulk shift approaches to generate regional seabed maps: Bay of fundy, canada," *Journal of Maps*, vol. 19, 12 2023. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/17445647.2023.2223629

[22] B. R. Wilson, C. J. Brown, J. A. Sameoto, M. Lacharité, A. M. Redden, and V. Gazzola, "Nc-nd license mapping seafloor habitats in the bay of fundy to assess megafaunal assemblages associated with modiolus modiolus beds," *Estuarine, Coastal and Shelf Science*, 2021. [Online]. Available: https://doi.org/10.1016/j.ecss.2021.107294

[23] F. Althaus, N. Hill, R. Ferrari, L. Edwards, R. Przeslawski, C. H. Schönberg, R. Stuart-Smith, N. Barrett, G. Edgar, J. Colquhoun, M. Tran, A. Jordan, T. Rees, and K. Gowlett-Holmes, "A standardised vocabulary for identifying benthic biota and substrata from underwater imagery: The catami classification scheme," *PLOS ONE*, vol. 10, p. e0141039, 10 2015. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0141039

[24] I. Xu, B. Misiuk, S. C. Lowe, M. Gillis, C. J. Brown, and T. Trappenberg, "Hierarchical multi-label classification with missing information for benthic habitat imagery," *International Joint Conference on Neural Networks*, pp. 1–10, 9 2024. [Online]. Available: http://arxiv.org/abs/2409.06618

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-December, pp. 770–778, 12 2016. [Online]. Available: https://doi.org/10.1109/CVPR.2016.90

[26] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," pp. 12 310–12 320, 7 2021. [Online]. Available: https://proceedings.mlr.press/v139/zbontar21a.html

[27] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR 2021 - 9th International Conference on Learning Representations*, 10 2020. [Online]. Available: https://arxiv.org/abs/2010.11929v2

[28] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 9620–9629, 4 2021. [Online]. Available: https://arxiv.org/abs/2104.02057v4

[29] D. L. Hall and J. Llinas, "An introduction to multisensor data fusion," *Proceedings of the IEEE*, vol. 85, pp. 6–23, 1997. [Online]. Available: https://doi.org/10.1109/5.554205

[30] R. Bokade, A. Navato, R. Ouyang, X. Jin, C. A. Chou, S. Ostadabbas, and A. V. Mueller, "A cross-disciplinary comparison of multimodal data fusion approaches and applications: Accelerating learning through trans-disciplinary information sharing," *Expert Systems with Applications*, vol. 165, p. 113885, 3 2021. [Online]. Available: https://doi.org/10.1016/J.ESWA.2020.113885

[31] S. R. Stahlschmidt, B. Ulfenborg, and J. Synnergren, "Multimodal deep learning for biomedical data fusion: a review," *Briefings in bioinformatics*, vol. 23, 3 2022. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/35089332/

[32] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, pp. 457–468, 6 2016. [Online]. Available: https://arxiv.org/abs/1606.01847v3

[33] K. Yu, D. Conway, S. Berkovsky, J. Zhou, R. Taib, and F. Chen, "User trust dynamics: An investigation driven by differences in system performance," *International Conference on Intelligent User Interfaces, Proceedings IUI*, pp. 307–317, 3 2017. [Online]. Available: https://dl.acm.org/doi/10.1145/3025171.3025219

[34] J. Gao, P. Li, Z. Chen, J. Zhang, J. Gao, P. Li, Z. Chen, and J. Zhang, "A survey on deep learning for multimodal data fusion," *Neural Computation*, vol. 32, pp. 829–864, 5 2020. [Online]. Available: https://dx.doi.org/10.1162/neco_a_01273

[35] F. Castanedo, "A review of data fusion techniques," *The Scientific World Journal*, vol. 2013, p. 704504, 1 2013. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1155/2013/704504

[36] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, C. Sun, and G. Research, "Attention bottlenecks for multimodal fusion," *Advances in Neural Information Processing Systems*, vol. 34, pp. 14 200–14 213, 12 2021. [Online]. Available: https://doi.org/10.48550/arXiv.2107.00135

[37] J. Huang, J. Tao, B. Liu, Z. Lian, and M. Niu, "Multimodal transformer fusion for continuous emotion recognition," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2020-May, pp. 3507–3511, 5 2020. [Online]. Available: https://doi.org/10.1109/ICASSP40776.2020.9053762

[38] W. Wang, D. Tran, and M. Feiszli, "What makes training multi-modal classification networks hard?" *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 12 692–12 702, 5 2019. [Online]. Available: https://arxiv.org/abs/1905.12681v5

[39] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, 9 2017. [Online]. Available: https://doi.org/10.1016/J.INFFUS.2017.02.003

[40] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE Signal Processing Magazine*, vol. 34, pp. 96–108, 11 2017. [Online]. Available: https://doi.org/10.1109/MSP.2017.2738401

[41] A. D. Snow, J. Whitaker, M. Cochran, I. Miara, J. V. den Bossche, C. Mayo, G. Lucas, P. Cochrane, J. de Kloe, C. Karney, J. J. Shaw, Filipe, G. Ouzounoudis, J. Dearing, G. Lostis, D. Hoese, B. Couwenberg, B. Jurd, C. Gohlke, C. Schneck, D. McDonald, M. Itkin, R. May, H. de Bittencourt, B. Little, P. S. Rahul, N. Eubank, Neil, and M. Taves, "pyproj4/pyproj: 3.7.0 release," Oct. 2024. [Online]. Available: https://doi.org/10.5281/zenodo.13864781

[42] S. Gillies *et al.*, "Rasterio: geospatial raster i/o for Python programmers," Mapbox, 2013–. [Online]. Available: https://github.com/mapbox/rasterio

[43] E. J. Gregrid, D. R. Haggarty, S. C. Davies, C. Fields, and J. Lessard, "Comprehensive marine substrate classification applied to canada's pacific shelf," *Public Library of Science*, 2021. [Online]. Available: https://doi.org/10.1371/journal.pone.0259156

[44] G. Fader, "Physiography, geography and bathymetry of the digby neck area, bay of fundy," 2005. [Online]. Available: https://novascotia.ca/nse/ea/whitespointquarry/09.Reference.Documents/15.Fader.Physiography.Geography.and.Bathymetry.pdf

[45] M. Lin, Q. Chen, and S. Yan, "Network in network," *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*, 12 2013. [Online]. Available: https://arxiv.org/abs/1312.4400v3

[46] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in *2015 IEEE Information Theory Workshop (ITW)*, 2015, pp. 1–5. [Online]. Available: https://doi.org/10.1109/ITW.2015.7133169

[47] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Weinberger, "Deep networks with stochastic depth," 2016. [Online]. Available: https://arxiv.org/abs/1603.09382

[48] G. Larsson, M. Maire, and G. Shakhnarovich, "Fractalnet: Ultra-deep neural networks without residuals," *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, 5 2016. [Online]. Available: https://arxiv.org/abs/1605.07648v4

[49] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2623–2631, 7 2019. [Online]. Available: https://arxiv.org/abs/1907.10902v1

[50] "optuna.samplers.tpesampler — optuna 4.2.0 documentation." [Online]. Available: https://optuna.readthedocs.io/en/v4.2.0/_modules/optuna/samplers/_tpe/sampler.html

[51] X. Ji, B. Yang, and Q. Tang, "Seabed sediment classification using multibeam backscatter data based on the selecting optimal random forest model," *Applied Acoustics*, vol. 167, p. 107387, 10 2020. [Online]. Available: https://doi.org/10.1016/J.APACOUST.2020.107387

[52] S. S. Hansen, V. B. Ernstsen, M. S. Andersen, Z. Al-Hamdani, R. Baran, M. Niederwieser, F. Steinbacher, and A. Kroon, "Classification of boulders in coastal environments using random forest machine learning on topo-bathymetric lidar data," *Remote Sensing 2021, Vol. 13, Page 4101*, vol. 13, p. 4101, 10 2021. [Online]. Available: https://www.mdpi.com/2072-4292/13/20/4101

[53] W. Xu, H. Cheng, S. Zheng, and H. Hu, "Predicted mapping of seabed sediments based on mbes backscatter and bathymetric data: A case study in joseph bonaparte gulf, australia, using random forest decision tree," *Journal of Marine Science and Engineering 2021, Vol. 9, Page 947*, vol. 9, p. 947, 8 2021. [Online]. Available: https://www.mdpi.com/2077-1312/9/9/947

[54] M. L. Zhang and Z. H. Zhou, "A review on multi-label learning algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, pp. 1819–1837, 2014. [Online]. Available: https://doi.org/10.1109/TKDE.2013.39

[55] R. C. Hasan, D. Ierodiaconou, L. Laurenson, and A. Schimel, "Integrating multibeam backscatter angular response, mosaic and bathymetry data for benthic habitat mapping," *PLOS ONE*, vol. 9, p. e97339, 5 2014. [Online]. Available: https://doi.org/10.1371/JOURNAL.PONE.0097339

[56] P. Xu, X. Zhu, and D. A. Clifton, "Multimodal learning with transformers: A survey," *IEEE Transactions on Pattern Analysis Machine Intelligence*, 2023. [Online]. Available: https://doi.org/10.1109/TPAMI.2023.3275156

Appendix A

Tables

The appendix contains the tables of results from the experiments. The captions of the table are describing the contents of the tables. The bold values suggest that the model is 2-σ better as compared to the other model at the given CATAMI hierarchical level.

CATAMI Hierarchy	Input Type	F1	Balanced Accuracy
		Backscatter	
Level 1	**Single**	0.948 ± 0.002	0.610 ± 0.016
	Grid	**0.961 ± 0.006**	**0.791 ± 0.035**
Level 2	**Single**	**0.907 ± 0.004**	0.370 ± 0.020
	Grid	0.892 ± 0.004	**0.614 ± 0.041**
Level 3	**Single**	0.810 ± 0.017	0.767 ± 0.027
	Grid	**0.860 ± 0.029**	**0.814 ± 0.036**
		Bathymetry	
Level 1	**Single**	0.932 ± 0.005	0.574 ± 0.005
	Grid	**0.962 ± 0.005**	**0.704 ± 0.056**
Level 2	**Single**	0.676 ± 0.004	0.376 ± 0.021
	Grid	**0.833 ± 0.038**	**0.521 ± 0.014**
Level 3	**Single**	0.767 ± 0.011	0.735 ± 0.016
	Grid	**0.857 ± 0.007**	**0.816 ± 0.009**

Table A.1: The performance metric scores for Random Forest models across the different CATAMI hierarchy levels for Single-point and context-based inputs.

Model Name	CATAMI Hierarchy	F1	Balanced Accuracy
SMR	**Level 1**	0.893 ± 0.008	0.532 ± 0.006
	Level 2	**0.743 ± 0.048**	**0.530 ± 0.037**
	Level 3	**0.967 ± 0.012**	**0.933 ± 0.030**
DMR	**Level 1**	0.907 ± 0.024	0.572 ± 0.053
	Level 2	0.625 ± 0.033	0.383 ± 0.059
	Level 3	0.813 ± 0.011	0.684 ± 0.027

Table A.2: Performance metrics of the dual-encoder data fusion models using SMR and DMR for sonar data as mentioned in Subsection 3.3.2. The two modalities used here are images and backscatter.

Model Name	CATAMI Hierarchy	F1 ± Std	Balanced Accuracy ± Std
SMR	**Level 1**	0.913 ± 0.004	**0.713 ± 0.016**
	Level 2	**0.616 ± 0.005**	0.355 ± 0.009
	Level 3	**0.966 ± 0.016**	**0.929 ± 0.034**
DMR	**Level 1**	0.912 ± 0.004	0.541 ± 0.003
	Level 2	0.568 ± 0.074	0.341 ± 0.088
	Level 3	0.886 ± 0.018	0.800 ± 0.032

Table A.3: Performance metrics of the dual-encoder data fusion models using SMR and DMR for sonar data as mentioned in Subsection 3.3.2. The two modalities used here are images and bathymetry.

Model Name	CATAMI Hierarchy	F1	Balanced Accuracy
bsSMR-bmSMR	**Level 1**	0.925 ± 0.019	0.717 ± 0.041
	Level 2	0.682 ± 0.033	0.466 ± 0.074
	Level 3	0.976 ± 0.005	0.957 ± 0.005
bsDMR-bmDMR	**Level 1**	0.939 ± 0.018	0.739 ± 0.035
	Level 2	0.659 ± 0.018	0.503 ± 0.064
	Level 3	0.981 ± 0.004	0.953 ± 0.006
bsSMR-bmDMR	**Level 1**	0.892 ± 0.022	0.539 ± 0.015
	Level 2	0.734 ± 0.038	**0.621 ± 0.019**
	Level 3	0.974 ± 0.004	0.946 ± 0.010
bsDMR-bmSMR	**Level 1**	0.910 ± 0.031	0.680 ± 0.063
	Level 2	0.718 ± 0.019	0.558 ± 0.059
	Level 3	0.948 ± 0.013	0.892 ± 0.020

Table A.4: Comparison of triple-encoder fusion models from Subsection 3.3.3 using different combinations of shallow (SMR) and deep (DMR) marginal representations for backscatter (bs) and bathymetry (bm) data.

Model	CATAMI Hierarchy	F1	Balanced Accuracy
ResNet50	**Level 1**	**0.957 ± 0.002**	**0.711 ± 0.032**
	Level 2	0.703 ± 0.003	**0.561 ± 0.017**
	Level 3	**0.986 ± 0.004**	**0.980 ± 0.002**
ViT	**Level 1**	0.951 ± 0.006	0.610 ± 0.008
	Level 2	0.700 ± 0.002	0.496 ± 0.016
	Level 3	0.945 ± 0.002	0.513 ± 0.010

Table A.5: Performance metrics of the baseline ResNet50 and ViT models (fine-tuned only on Image data) for all the levels of the CATAMI hierarchy.

Model	CATAMI Hierarchy	F1	Balanced Accuracy
		Backscatter	
ResNet 50	**Level 1**	0.962 ± 0.002	0.652 ± 0.015
	Level 2	0.745 ± 0.046	0.540 ± 0.024
	Level 3	0.994 ± 0.000	0.988 ± 0.000
ViT	**Level 1**	0.971 ± 0.006	**0.853 ± 0.046**
	Level 2	**0.919 ± 0.025**	**0.685 ± 0.025**
	Level 3	0.995 ± 0.003	0.991 ± 0.005
		Bathymetry	
ResNet 50	**Level 1**	0.960 ± 0.001	0.649 ± 0.016
	Level 2	0.714 ± 0.030	0.576 ± 0.033
	Level 3	0.994 ± 0.000	0.988 ± 0.000
ViT	**Level 1**	0.958 ± 0.006	**0.762 ± 0.040**
	Level 2	**0.917 ± 0.022**	**0.710 ± 0.023**
	Level 3	0.997 ± 0.002	0.991 ± 0.005

Table A.6: Performance metrics of the ResNet50 and ViT-based dual-encoder fusion models for both backscatter and bathymetry.

Model	CATAMI Hierarchy	F1	Balanced Accuracy
ResNet 50	**Level 1**	0.966 ± 0.003	0.643 ± 0.018
	Level 2	0.707 ± 0.003	0.582 ± 0.022
	Level 3	0.994 ± 0.000	0.988 ± 0.000
ViT	**Level 1**	0.962 ± 0.006	**0.827 ± 0.049**
	Level 2	**0.936 ± 0.025**	**0.748 ± 0.030**
	Level 3	0.993 ± 0.004	0.988 ± 0.007

Table A.7: Performance metrics of the ResNet50 and ViT-based triple-encoder fusion models at different levels of the CATAMI hierarchy.

CATAMI Hierarchy	Backscatter			Bathymetry		
	Level 1	**Level 2**	**Level 3**	**Level 1**	**Level 2**	**Level 3**
F1	0.955	0.685	0.793	0.824	0.590	0.894
Balanced Accuracy	0.612	0.368	0.711	0.462	0.393	0.828

Table A.8: The performance metric of the baseline BayesianKDE model from Experiment 1 (Subsection 3.3.1).

Model	CATAMI Hierarchy	Image BA	Image F1	Sonar BA	Sonar F1
		Backscatter			
ResNet50	Level 1	0.648 ± 0.015	0.959 ± 0.002	0.333 ± 0.000	0.495 ± 0.000
	Level 2	0.538 ± 0.029	0.730 ± 0.049	0.200 ± 0.000	0.422 ± 0.000
	Level 3	**0.988 ± 0.000**	0.994 ± 0.000	0.500 ± 0.000	0.571 ± 0.000
ViT	Level 1	**0.791 ± 0.082**	**0.967 ± 0.005**	**0.362 ± 0.021**	**0.845 ± 0.007**
	Level 2	**0.658 ± 0.030**	**0.919 ± 0.019**	0.197 ± 0.006	0.417 ± 0.009
	Level 3	0.981 ± 0.006	0.992 ± 0.003	0.500 ± 0.000	0.571 ± 0.000
		Bathymetry			
ResNet50	Level 1	0.632 ± 0.020	0.959 ± 0.001	0.333 ± 0.000	0.495 ± 0.000
	Level 2	0.499 ± 0.018	0.693 ± 0.004	0.200 ± 0.000	**0.422 ± 0.000**
	Level 3	0.988 ± 0.000	0.994 ± 0.000	0.500 ± 0.000	0.571 ± 0.000
ViT	Level 1	**0.762 ± 0.040**	0.958 ± 0.006	0.333 ± 0.000	0.495 ± 0.000
	Level 2	**0.596 ± 0.016**	**0.907 ± 0.022**	0.208 ± 0.007	0.418 ± 0.001
	Level 3	0.986 ± 0.005	0.995 ± 0.001	0.500 ± 0.000	0.571 ± 0.000

Table A.9: Performance Metrics for Image and sonar encoders of the ResNet50 and ViT-based dual-encoder models across CATAMI hierarchical levels for Backscatter and Bathymetry data.

Metric	Level 1	Level 2	Level 3
		ResNet50	
Image BA	0.620 ± 0.003	0.493 ± 0.012	0.986 ± 0.005
Image F1	**0.961 ± 0.002**	0.697 ± 0.003	0.992 ± 0.003
Backscatter BA	0.333 ± 0.000	0.200 ± 0.000	0.500 ± 0.000
Backscatter F1	0.495 ± 0.000	**0.422 ± 0.000**	0.571 ± 0.000
Bathymetry BA	0.333 ± 0.000	0.200 ± 0.000	0.500 ± 0.000
Bathymetry F1	0.495 ± 0.000	0.422 ± 0.000	0.571 ± 0.000
		ViT	
Image BA	**0.708 ± 0.049**	**0.567 ± 0.030**	0.988 ± 0.007
Image F1	0.956 ± 0.006	**0.920 ± 0.031**	0.993 ± 0.004
Backscatter BA	0.371 ± 0.071	0.200 ± 0.000	0.500 ± 0.000
Backscatter F1	0.581 ± 0.162	0.416 ± 0.010	0.571 ± 0.000
Bathymetry BA	0.342 ± 0.017	**0.314 ± 0.010**	**0.742 ± 0.017**
Bathymetry F1	0.563 ± 0.136	**0.529 ± 0.065**	**0.856 ± 0.015**

Table A.10: Performance Metrics for Image and sonar encoders of the ResNet50 and ViT-based triple-encoder models across CATAMI hierarchical levels.