AN IMPROVED CLASSIFICATION OF UNDERWATER SHIP-ENGINE AUDIOS USING SIAMESE NETWORK

By
MARYAM SAJID

NATIONAL UNIVERSITY OF MODERN LANGUAGES
ISLAMABAD
August, 2024

AN IMPROVED CLASSIFICATION OF UNDERWATER SHIP-ENGINE AUDIOS USING SIAMESE NETWORK

By

MARYAM SAJID

BSIT, International Islamic University, Islamabad, 2020

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

In **Computer Science**

To

FACULTY OF ENGINEERING & COMPUTER SCIENCE

NATIONAL UNIVERSITY OF MODERN LANGUAGES
ISLAMABAD

NATIONAL UNIVERSITY OF MODERN LANGUAGES FACUTY OF ENGINEERING & COMPUTING

THESIS AND DEFENSE APPROVAL FORM

The undersigned certify that they have read the following thesis, examined the defense, are satisfied with overall exam performance, and recommend the thesis to the Faculty of Engineering and Computing for acceptance

Thesis Title: An Improved Classification of Underwater Ship-Engine Audios Using Siamese Network

Submitted By: Maryam Sajid **Registration #:**64 MS/CS/F21

Master of Science in Computer Science (MSCS)
Degree Name in Full

Computer Science
Name of Discipline

Dr. Quara-tul-Ain Safdar
Research Supervisor

Signature of Research Supervisor

Dr. Hina Ashraf
Research Co-Supervisor

for
Signature of Research Co-Supervisor

Dr. Fazli Subhan
Head of Department (CS)

Signature of HoD (CS)

Dr, M. Noman Malik
Name of Dean (FE & C)

Signature of Dean (FE & C)

August 30th, 2024

AUTHOR'S DECLARATION

I Maryam Sajid

Daughter of Sajid Habib Qureshi

Registration # 64 MS/CS/S21

Discipline Computer Science

Candidate of **Master of Science in Computer Science (MSCS)** at the National University of Modern Languages do hereby declare that the thesis **An Improved Classification of Underwater Ship-Engine Audios Using Siamese Network** submitted by me in partial fulfillment of MSCS degree, is my original work, and has not been submitted or published earlier. I also solemnly declare that it shall not, in future, be submitted by me for obtaining any other degree from this or any other university or institution. I also understand that if evidence of plagiarism is found in my thesis/dissertation at any stage, even after the award of a degree, the work may be cancelled and the degree revoked.

Signature of Candidate

Maryam Sajid
Name of Candidate

August 30th, 2024
Date

ABSTRACT

Title: An Improved Classification of Underwater Ship-Engine Audios Using Siamese Network

Developing a reliable ship classification system using underwater acoustics is crucial due to limited labeled data, dynamic underwater conditions, and noise interference, reducing reliance on human sonar operators vulnerable to weather and fatigue. Improving underwater acoustic target classification requires addressing shortcomings in feature extraction, dataset availability, feature diversity, and classifier selection, but integrating multiple techniques must balance gains against increased costs, time, and system complexity. This study proposes a novel feature extraction technique that reduces the computational cost, complexity and increases the robustness of model. In this proposed technique, audios are segmented into chunks and spectrograms are calculated for them. The dataset is arranged in the form of triplets that are fed into the siamese network that are based on triplet loss, generates feature vectors. The goal of the siamese network is to learn an embedding space in which similar classes are grouped together and dissimilar classes are further separated. These extracted features may be fed into a classifier. Classifier will then classify the correct classes on the basis of given results. The model's performance is evaluated on Shipears dataset. Furthermore, accuracy, precision, recall, f1-score and ROC curve are used to evaluate the performance of popular classifiers, k-NN, SVM, RF, DT. Overall accuracy of our model reaches 96.4167% which reduces the complexity.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

MFCC	-	Mel Frequency Cepstral Coefficient
USCSN	-	Underwater Ship-Engine Classification Using Siamese Network
AUVs	-	Autonomous Underwater Vehicles
ROVs	-	Remotely Operated Vehicles
ML	-	Machine Learning
STFT	-	Short-Time Fourier Transform
LOFAR	-	Low-Frequency Analysis and Recording
GAN	-	Generative Adversarial Networks
SVM	-	Support Vector Machines
k-NN	-	K Nearest Neighbor
USAR	-	Underwater Ship-Engine Acoustic Recognition
SNE	-	t-Distributed Stochastic Neighbor Embedding
ROC	-	Receiver Operating Characteristic
CNN	-	Conventional Neural Network
ReLU	-	Rectified Linear Unit
LM	-	Long Mel
CCTZ	-	Chroma, Spectral Contrast, Tonnetz, And Zero-Cross Ratio
PCA	-	Principal Component Analysis
ESC	-	Environment Sound Classification
EM	-	Expectation Maximization
DNN	-	Deep Neural Networks
LSTM	-	Long Short-Term Memory
DML	-	Deep Mutual Learning
SCAE	-	Separable Convolution Autoencoder
ONC	-	Ocean Networks Canada
RBM	-	Restricted Boltzmann Machine
UATC-DENSENET	-	Underwater Acoustic Target Classification Densenet

LAN	-	Local Area Network
ABNN	-	Attention-Based Neural Network
AIS	-	Automatic Identification System
DCGAN	-	Deep Convolutional Generative Adversarial Network
UATR	-	Underwater Acoustic Target Recognition
MWSA	-	Multi-Window Spectral Analysis
TP	-	True Positive
TN	-	True Negative
FP	-	False Positive
FN	-	False Negative
WND	-	Wigner-Ville Distribution
DT	-	Decision Tree
RF	-	Random Forest

ACKNOWLEDGMENT

“All praise to Allah, the lord of the worlds, and His Prophet Muhammad (peace be upon him), his family and his companions.”

First of all, I wish to express my gratitude and deep appreciation to Almighty Allah, who made this study possible and successful.

This study would not be accomplished unless the honest espousal that was extended from several sources for which I would like to express my sincere thankfulness and gratitude. Yet, there were significant contributors for my attained success and I cannot forget their input, especially my research supervisors, Asst. Prof. Dr. Quara-tul-Ain Safdar, who did not leave any stone unturned to guide me during my research journey.

I shall also acknowledge the extended assistance from the administrations of Department of Computer Sciences who supported me all through my research experience and simplified the challenges I faced. For all whom I did not mention but I shall not neglect their significant contribution, thanks for everything.

DEDICATION

This thesis work is dedicated to my parents and my teachers throughout my education career who have not only loved me unconditionally but whose good examples have taught me to work hard for the things that I aspire to achieve.

CHAPTER 1

INTRODUCTION

1.1 Overview

Underwater acoustic target recognition has recently drawn attention from scientific and technical experts as a result of the advancements in science and technology. Recognizing underwater acoustic targets is a very becoming essential. Underwater acoustic target detection is challenging than traditional speech recognition due to the distorted radiated noise caused by the complex underwater habitat. The gradual progress of technical improvement is driving the demand for robust underwater acoustic target detection methods. [1-3]. Underwater acoustic ship-engine classification is often carried out by skilled sonar operators, who work long hours and are susceptible to weather changes. Therefore, it is important to create a reliable recognizing system to do the task now done by people in identifying ship-radiated noise [4].

Through local connections and weight sharing, deep learning offers an effective method for classifying targets in the realm of image processing. It involves the design of classifiers and feature extraction. Deep learning might increase classification accuracy and efficiency by avoiding feature loss and dimension catastrophe in comparison to conventional approaches. The typical approach would be to convert audio data into image data when using deep learning to classify underwater targets. The target is then classified using the pre-processed image data that was given to the classifier [2].

The basic structure for classification of underwater ship-engine audios is shown in figure 1.1. First, the ship-radiated noise is given as input, then features are extracted through some feature extraction techniques. Data augmentation technique may be used to generate fake samples in case of limited dataset. The classifier with the help of extracted features, classify the true classes for ship-engines.

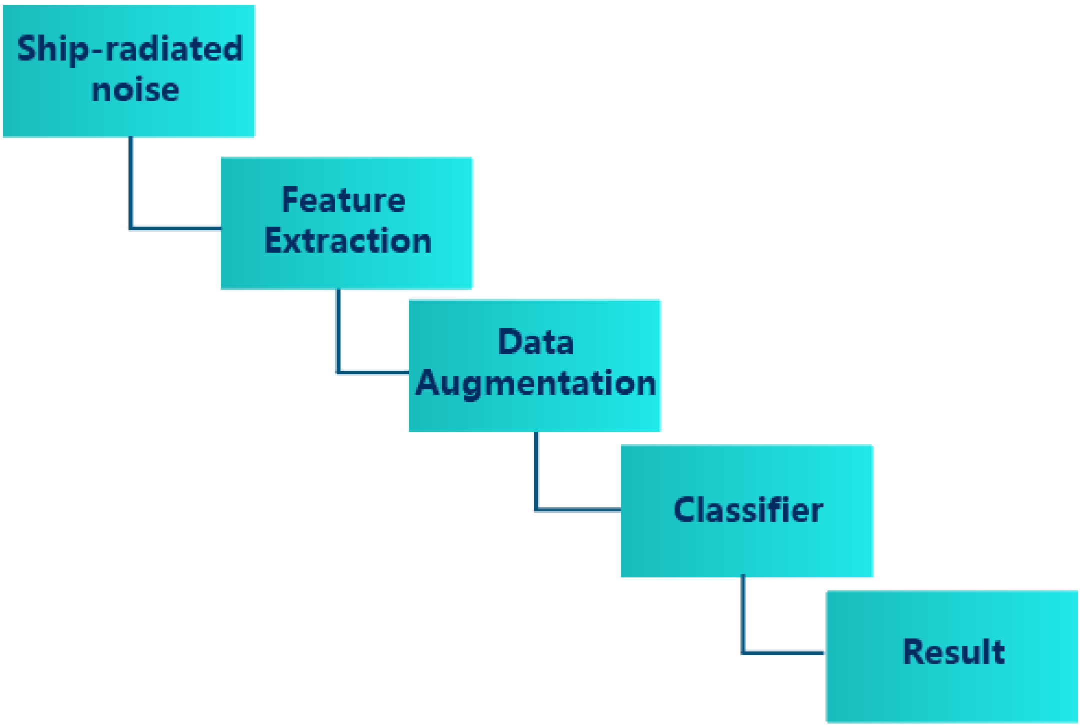

Figure 1.1: Basic structure of classification of underwater ship-engine

Two important steps from the above figure are: [3] first, the hydroacoustic signals are processed using techniques to extract interesting features. Classification algorithms (classifiers) are then used using the recovered features as input, in order to classify the ship-engine. Scholars have put through comprehensive research on ship-engine classification deep learning methods and the research mainly emphasis on feature extraction module. Feature extraction is a process of either removing redundant or extracting only the relevant information from original data to achieve dimensionality reduction.

The improvements are continuously made to the feature extraction and classifier training processes to increase classification accuracy [4]. Underwater audio data is limited, due to security reasons or expensive to record in terms of resources. Many scholars have used publicly available dataset, and some of them have used their own dataset which they donot share. There are two ways to overcome this problem; data augmentation or, diversify the features. Former technique generates fake samples from the existing limited samples while latter diversify the features enough so that the classifier could easily and accurately classify the dataset.

To increase classification accuracy, our strategy employs Siamese vector and multimodal feature extraction methods. To see how approaches affect the model, we will use the technique to diversity the characteristics rather than data augmentation. In order to examine the impact of classification accuracy on underwater data, we will be using different popular classification algorithms. We'll use the accessible ShipEars dataset to evaluate the model's performance. A training set was created by randomly selecting 80% of the feature samples from the recorded audios, and a testing set was created by randomly selecting 20% of the feature samples from the training set.

1.2 Motivation

Underwater classification of ship-engine is performed by domain specialists [6]. Long-time work and weather conditions. More accurate underwater acoustic classification methods must be investigated. Deep learning methods are created for assistance of human experts. For example, monitors or tests to assist doctors. Underwater audio dataset is limited, usually expensive or not available due to security [5]. There are two ways to overcome this problem:

1.2.1 Data Augmentation

Neural network training demands a large amount of data. Learned networks exhibit poor generalization and underdetermination of parameters in the low-data environment. This is lessened via data augmentation, which makes better use of already-existing data [22]. Data augmentation is used to create new data points from preexisting data in order to fictitiously increase the volume of data. This includes adding modest adjustments to data or creating new data points with deep learning models. In order to improve the model's ability to generalize to new cases, it is intended to produce variants of the original data that yet reflect the same underlying patterns. For example, data augmentation comes in handy when working with little amounts of labelled training data.

1.2.2 Diversify Features

Building strong machine learning models requires diversifying its features, which aids in the model's good generalization to various patterns found in the data. The following techniques aim to diversify features in order to achieve precise class identification:

i) Engineering Features: Higher-order polynomial features can be introduced to capture non-linear relationships. Terms of interaction means, merge two or more characteristics to capture their combined impact. Moreover, binding or discretization means to identify non-linear patterns, divide continuous features into discrete bins.
ii) Scaling Features: Preprocessing methods like normalization and standardization are frequently employed to get numerical features ready for machine learning models. In order to guarantee that every numerical feature contributes equally to the model, normalization scales the features to a standard range, usually between 0 and 1. However, standardization makes the features comparable by changing their mean and standard deviation to zero and one, which helps with gradient descent optimization.
iii) Managing Data Inequalities: Consider using techniques like oversampling the minority class or under sampling the majority class, especially if the distribution of classes is unequal, to guarantee that each class in dataset has enough exposure to samples. Under

sampling refers to lowering the number of samples in the majority class, whereas oversampling implies creating artificial samples of the minority class in order to balance the class distribution. In order to identify trends or patterns throughout time, temporal aspects must also be included when working with time-series data.

1.2.3 Constraints of Underwater Acoustic Target Recognition (UATR)

Underwater ship-engine audio classification comes with its own set of difficulties and limitations. It takes a combination of feature engineering, machine learning, and signal processing methods to overcome these limitations. When developing and accessing classification models, it's critical to take into account the particulars of the underwater environment as well as the distinctive qualities of ship-engine audios. To overcome these obstacles, cooperation with subject matter experts and having access to a variety of representative datasets are also essential. The following are some typical constraints related to the categorization of underwater ship-engine audios:

i) Background noise: Ship-engine audios can be impacted by ambient noise from nearby sources, such as other ships, marine life, or underwater currents, much like other underwater acoustic applications. Moreover, it can be difficult to distinguish the target ship's engine noises from background noise, which reduces classification accuracy.
ii) Variability in Engine Designs and Ship Types: Engine designs and operational parameters vary throughout ship types. However, it may be more challenging to develop a universal categorization model that is effective in a variety of marine conditions due to variations in ship types and engine configurations.
iii) Changes in Depth and Distance: Depth, distance, and the acoustic characteristics of the water all affect how sound travels underwater. However, depth and distance from the recording sensor can affect the strength and frequency of ship-engine noises, making it difficult to classify them consistently.
iv) Insufficient Training Information: Because of the limited access to active ships and regulated testing facilities, gathering labelled training data for ship-engine audios might

be difficult. Accurate and reliable classification model training may be hampered by a lack of training data.

v) Temporal Dynamics: The noises produced by ship engines can have temporal dynamics, changing in character as the engine runs through various stages (such as starting, cruising, and idling). Accurate categorization depends on capturing and comprehending the temporal dynamics, and models must take these variances into account.

vi) Conditions with Hydroacoustic: Underwater sound transmission can be affected by hydroacoustic factors like pressure, temperature, and salt of the water. However, the changes in hydroacoustic circumstances may have an impact on the features and quality of audio recordings of ship engines.

vii) Deployment Challenges for Sensors: It can be difficult to deploy sensors in the undersea environment, particularly in places where there is a lot of shipping traffic. Inadequate sensor deployment could lead to coverage gaps in monitoring, which would reduce the classification system's overall efficacy.

1.2.4 Applications of Underwater Acoustic Target Recognition (UATR)

The applications demonstrate the variety of ways that underwater ship-engine audios can be applied for real-world issues, such as scientific research, marine safety, and environmental preservation.

i. Environmental Surveillance: Investigating the sounds that ships make underwater can reveal important information about the ways that maritime activities affect marine ecosystems, which include habitats and marine life. One type of underwater noise pollution that is particularly important in this regard is ship-engine noise. Marine biodiversity and ecosystems can be conserved and protected by effectively managing and controlling the environmental impact of ship-engine noise pollution by monitoring and knowledge of its consequences.

ii. Underwater Cartography and Navigation: Underwater navigation systems can be improved by ship-engine audio data, which gives autonomous underwater vehicles (AUVs) and remotely operated vehicles (ROVs) more information. Identification and characterization of underwater pathways and structures are made possible by the analysis of ship-engine noises in underwater route and infrastructure mapping. The efficiency and effectiveness of marine activities are improved by this data, which makes it easier to navigate and explore underwater habitats with greater accuracy.

iii. Marine Bioacoustics and Marine Biology: Bioacoustics research uses audio data from ship engines to identify and analyses marine creatures. By identifying and examining marine species based on their acoustic traits, researchers can learn more about their behavior and communication. By tracking how marine animals and other species respond to ship noises, behavioral studies may be conducted, which improves our knowledge of marine ecosystems and helps conservation efforts.

iv. Monitoring of Underwater Infrastructure: The monitoring of underwater infrastructure, such as cables and pipelines, to spot possible issues or damage, is made possible via ship-engine audio data. One can monitor construction operations, including dredging and other underwater engineering projects, by listening for ship-engine noises. This promotes efficient management of marine building operations, guarantees the integrity and safety of submerged infrastructure, and makes it easier to identify problems in a timely manner.

v. Defense and Security: Audio monitoring of ship engines can be employed for security purposes, assisting in the identification and detection of vessels entering sensitive or prohibited regions. Moreover, ship-engine audio monitoring helps to improve maritime security protocols.

vi. Scientific Investigations: The sounds produced by ship engines can be used to gather important information about the properties of various ocean locations and the study of underwater acoustics. Tracking the audio from ship engines can help determine how marine activities affect the climate.

vii. Monitoring of Maritime Traffic: In crowded maritime locations, ship-engine sound analysis can be utilized to identify and track vessels, aiding in efficient traffic management and monitoring. Moreover, by giving real-time information on the movements of other vessels, the use of ship-engine audios aids in the prevention of collisions between them.

1.3 Problem Background

Understanding the underwater environment and all of its facets is essential to preserving and improving it. A crucial component of this information has to do with sound, whether it comes from natural or artificial sources. When researching maritime biodiversity and conducting vessel monitoring, the challenge of recognizing and classifying pertinent sounds is important.

Underwater sound recognition, thus, is a useful adjunct to existing ocean monitoring methods that rely on the classification of underwater images [1-3]. Researchers have worked on several kinds of projects over the years to identify ship-engine's sounds. Sonars and operators skilled in identifying different kinds of vessels from sonar echo signals were first used for these tasks. However, since it's a low automated task that requires the operator to concentrate on a specific task, human error can always occur. Passive hydrophone-based solutions are chosen for sustainability considerations, as active sonar-like systems are now known to pose a threat to various fish and cetacean species. This is the reason why underwater noise signals captured by passive hydrophone devices are increasingly being recognized using machine learning (ML)-based techniques.

Usually, these methods are divided into two stages. In order to categorize or identify certain targets, first the hydroacoustic signals are processed in order to extract features of interest, a process known as feature extraction. Afterwards, classification algorithms, or classifiers, are used using the retrieved features as input. Figure 1.2 shows the overall flow of underwater classification.

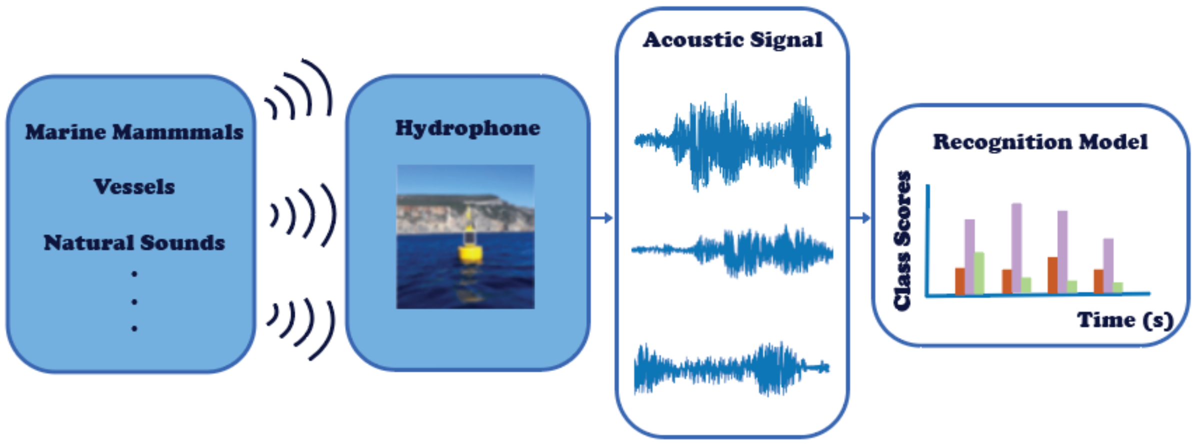

Figure 1.2: Marine acoustic signature recognition [3]

The feature extraction technique, sample set management, and classifier design are important areas of study in undersea classification research. Power spectrum analysis works well as a feature extraction technique for ship-radiated noise because of its short-time stability. An intricate distribution in the time domain of signal energy is transformed into a comparatively straightforward distribution in the frequency domain by power spectrum analysis. As a reliable feature for classification, the power spectrum represents the ship's radiated noise signal. For the classifier to successfully classify ship targets, the power spectrum is employed as an input [11]. Additionally, the use of the short-time Fourier transform (STFT), also known as low-frequency analysis and recording (LOFAR), it is possible to obtain the spectrogram of ship radiated noise signals. To categorize civil ships, large ships, and ferries, researchers [23] developed a deep learning recognition technique based on time-domain data and the LOFAR spectrum. Underwater acoustic signal characteristics are frequently extracted using wavelet analysis, allowing for the acquisition of energy distributions with various time-frequency resolutions inside a single spectrogram. Wavelet analysis was utilized to extract features from submerged audio sources. But the aforementioned feature extraction approach lacks flexibility in terms of handling various feature kinds since it uses a model with a set of parameters to extract the desired features.

Moreover, In the previous studies, the training set was randomly selected from 80% of the feature samples. The training samples comprised records from all dataset. Even yet, the classifier will never see the sample it obtained from this record and may encounter recorded data of new vessels when it is actually in operation [5].

However, underwater dataset is limited and needs an approach to handle it. Researchers have worked on different methods related to small dataset. GAN, or generative adversarial networks, are an effective method to get around the problems that arise with working with small datasets. The generator and discriminator neural networks in a GAN are simultaneously taught through advertising. Without being aware of the target data, the erstwhile blind forger attempts to create samples from a low-dimensional latent vector to a high-dimensional vector. The latter, an investigator, determines if the generator's output is genuine or fraudulent. Following training, the generator may create believable examples that resemble the target samples [25]. Generally speaking, GANs are unsupervised learning algorithms made to produce artificial data. They lack a mechanism by default for integrating labelled data, which is necessary for classification jobs. Classification performance may be impacted by biassed or inadequate representations of the classes as a result. GAN's intrinsic instability makes them difficult to train. Instable training has the potential to impair model convergence and reduce classification efficacy. GANs might have trouble with unequal class. However, to address these problems, there are more methods that deal with small dataset which are not explored in domain of underwater classification.

1.4 Problem Statement

Underwater audio data is limited and not available due to security reasons. The scarcity of data used for classification can be addressed by either data augmentation or by extracting feature that are unique and diverse enough to recognize unseen data. Moreover, the feature extraction approach uses a model with a set of parameters to extract the necessary characteristics, it is not flexible enough to handle different kinds of features. Moreover, human mistake can always happen because it's a low automated operation that needs the operator to focus on a particular task [2-4].

1.5 Research Questions

i. What is the impact of using features only from Siamese network on accuracy?
ii. What is the impact of applying different classification methods on Siamese network?

1.6 Aim of Research

This study aims to use the feature from Siamese networks to diversify feature extracted for recognition. Moreover, to enhance the accuracy and effectiveness of classifying underwater ship-engine audio signals using siamese networks compared to traditional classification methods or other deep learning architectures. Siamese networks operate well in learning scenarios that include one or few shots and have a small amount of labelled data available. The idea would be to use the siamese architecture to learn from tiny datasets efficiently, without requiring large numbers of labelled samples. The main objective would be to outperform current techniques in the classification of underwater ship-engine audio signals. Moreover, the siamese network approach's performance would probably be compared in the study to baseline techniques like conventional machine learning classifiers. Through the accomplishment of these goals, the study hopes to further the development of underwater acoustic signal processing and classification methods, which may find use in a range of marine-related fields and enterprises.

1.7 Research Objectives

i. To propose the modified USCSN (underwater ship-engine classification using siamese network) with Siamese based feature extraction technique.
ii. To evaluate the performance of the proposed model by incorporating different classification methods.

1.8 Scope of Research:

Underwater classification methods are developing due to particular challenges in marine environment. Underwater ship identification and classification is crucial for maritime security operations, which include spotting and tracking vessel movements in restricted or sensitive areas, enforcing maritime laws and regulations, and stopping illicit activities like piracy, smuggling, and illegal incursions.

Underwater ship classification systems are also used by port authorities and coastal security organizations to track marine traffic entering and leaving ports, spot suspicious or possibly dangerous vessels, and guarantee the protection of port infrastructure, ships, and workers. Furthermore, swift and precise ship classification can greatly increase the likelihood that search and rescue operations will be successful. Especially in delicate marine ecosystems and protected regions, underwater ship categorization is essential to environmental monitoring and preservation initiatives. Authorities can prevent pollution events, evaluate the possible effects of maritime operations on marine environments, and enforce environmental restrictions to minimize harm to marine biodiversity by identifying and tracking ships. Underwater ship classification also provides information about the makeup and behavior of maritime traffic, which is useful for managing marine resources effectively.

Encouraging scientific research into the undersea environment, preserving cultural legacy, safeguarding marine ecosystems, underwater ship classification, and sustainable marine resource management all depend on underwater ship classification. Authorities are able to maintain safe and responsible maritime operations while preserving the marine environment and its various ecosystems by precisely identifying and monitoring ships underwater.

1.9 Research Organization:

The thesis's remaining sections are arranged as follows:
Following a brief history of the subject,

Chapter 2 looks at analogous issues with underwater classification methods. A classification and discussion of the positive and negative aspects of the existing techniques are provided. A thorough operational working comparison of the several underwater network protocols is also given. Chapter 2 finally addresses the research gap that was utilized to create and refine the routing protocol.

Chapter 3 focuses on overall work of underwater classification method. It discusses the problems and solutions for the existing underwater techniques. The flow of the operational framework is explained in detail in Chapter 3. It provides information on the parameters utilized in the simulation of the proposed method as well as the dataset description, proposed method layers, and classifiers explanation.

Chapter 4 dealt with performance simulation studies of the proposed model. We investigate the effects of applying several metrics for performance, such as accuracy, confusion matrix. Additionally, a comparison of the classifiers is provided, along with an explanation and a number of figures. SNE-plot that describe the effect of accuracy on proposed model. However, the ROC curve testing explains the true positive rate over the thresholds.

The work's principal discoveries and contributions are outlined in Chapter 5, along with some possible future possibilities.

CHAPTER 2

LITERATURE REVIEW

2.1 Overview

Deep learning technology is currently in demand across numerous industries. It has made a new path for the underwater acoustic ship engine classification. An integral component of this understanding pertains to sound, regardless of its natural or artificial origins [2]. When researching maritime biodiversity and conducting vessel monitoring, the challenge of identifying and categorizing pertinent sounds is crucial.

As result, marine acoustic signature recognition enhances other ocean surveillance methods that rely on the classification of underwater images. It can even enhance maritime surface imagery from sources like drones or cameras. It is important to remember that underwater photography has a maximum depth of about fifty meters, yet sound waves in seawater can travel thousands of km.

Researchers have worked on a number of projects over the years to identify vessel sounds. Sonars and operators skilled in identifying different kinds of vessels from sonar echo signals were first used for these tasks. However, since it's a low automated task that requires the operator to concentrate on a specific task, human error can always occur. Passive hydrophone-based solutions are chosen for sustainability considerations, as active sonar-like systems are now known to pose a threat to various fish and cetacean species [3].

2.2 Underwater Recognition Techniques

Various technologies and approaches are used in underwater recognition techniques to recognize and classify objects, structures, or living organisms. These methods are essential in areas including oceanography, underwater archaeology, marine biology, and naval operations. Mostly, the scholars have used conventional neural network (CNN) to underwater acoustic ship engine classification [1].

2.2.1 Deep Learning Techniques

Vaz et al. [3] proposed CNN architecture with three convolutional layers and two fully connected layers. The underwater acoustic target detection framework was based on the combination of a CNN as the classifier and a mel spectrogram of the sound input with its first and second derivatives as features. Furthermore, this encourages the use of the method for real-time classification by employing brief analysis windows. First the audio signals were converted into spectrograms. The first and second derivatives of the mel spectrogram can be added to the feature extraction process to add details on the dynamic behavior of the parameters along the frequency axis for cepstral coefficients. Before being transformed to decibels, the derivatives were calculated from the difference between the mel spectrogram coefficients along the frequency axis. The mel spectrogram of each recording is then subjected to a rectangular window function with a defined length in frames, which divides the mel spectrogram into a group of windows of the same size.

The CNN architecture was used to classify the marine audios. Three convolutional layers and two fully connected layers make up the CNN architecture. 48 filters were used by the second and third convolutional layers, compared to 24 filters by the first. The zero-padding was equal to two, and the filter size is five by five. The stride of 1 was used by the filter to slice along the input. ReLU serves as an activation function in the first fully connected layer as well as the three convolutional layers. The SoftMax function was a common activation function in multiclass classification, is applied to the output layer. The number of classes is taken into consideration while sizing the fully connected output layer. The methodology was tested on

Ship-Ears dataset [5]. This dataset comprises of 90 acoustic recordings from 11 different ship types over a 15-to-10-minute period. They can be divided into four categories based on the types of ships, namely A, B, C, and D, and E for ambient noise, according to the annotation in the original dataset. Second dataset for generalization was marine dataset. Performance using Ship Ear's dataset was better with the feature combination of the mel spectrogram and the first and second derivatives than with the mel spectrogram alone. The two model's respective average accuracy was 88.8 percent and 83.2 percent. Second dataset was extended by using the data augmentation techniques. Three popular data augmentation techniques were used to artificially expand the dataset in order to balance the data on dolphins and humpback whales. Classification method CNN was also used for animal classification. The accuracy of animal classification was 78 percent.

Hong et al. [4] proposed three step feature extraction technique Long Mel (LM), Mel Frequency Cepstral Coefficient (MFCC), Chroma, Spectral Contrast, Tonnetz, and Zero-cross ratio (CCTZ). A lot of effort goes into trying to separate manually created features from ship-radiated noise and input them into various classifier types. One the one hand, Support Vector Machines (SVM) and Principal Component Analysis (PCA) techniques are commonly employed in the conventional machine learning feature extraction procedure. For instance, a strategy that uses SVM and the wave structure directly was proposed by researchers. Research was conducted on an extraction technique based on PCA and spectrum.

Two features that were frequently utilized in Environment Sound Classification (ESC) tasks with acceptable performance were those obtained from Mel filters of Mel Frequency Cepstral Coefficients and Log-Mel Spectrogram (LM). The influence of MFCC and its first-order differential MFCC or second-order MFCC features was demonstrated for underwater acoustic target detection, despite the fact that such characteristics originate from the speech or sound field. Furthermore, a large body of research suggests that the fusion feature can provide a more thorough depiction of ambient noises. Researchers utilizes the fusion feature of zero-crossing wavelength, peek-to-peek amplitude, and zero-crossing-wavelength difference for the identification of underwater sound targets. The baseline machine learning approach for ShipEar [5] might be a GMM-based classifier trained using the conventional expectation maximization (EM) technique; this approach has the best classification rate at 75.4 percent. Given the low success rate of machine learning-based approaches, deeper learning model-based approaches merit more investigation. An 84 percent accuracy rate was attained by a proposed feature optimization method that used Deep Neural Networks (DNN) and an optimizing loss function. For UATR, Deep Belief Nets were suggested. A UATR technique based on Restricted Boltzmann Machine achieves 93.17 percent accuracy on the ShipEars dataset [5]. The methodology was tested on Ship-Ears dataset [5]. This dataset comprises of 90 acoustic recordings from 11 different ship types over a 15- to 10-minute period. They can be divided into four categories based on the types of ships, namely A, B, C, and D, and E for ambient noise, according to the annotation in the original dataset. Warping the features, masking blocks of frequency channels, and masking blocks of time steps make up the augmentation policy. The ShipEar dataset's accuracy findings of 94.3 percent demonstrated that the suggested method achieved state-of-the-art accuracy.

Han et al. [2] proposed joint neural network that combines a large short-term memory network with a one-dimensional convolutional neural network. Deep neural networks automatically extract deep information to identify underwater audio targets. New methods were put out over time to increase the categorization accuracy of underwater sound targets. For the purpose of training a deep belief network, researchers employed a competitive learning process on the spectrum of ship radiated noise to improve the cluster performance. Based on the cepstrum, a convolutional neural network (CNN) was suggested for the simultaneous detection and range of broadband acoustic noise sources. Subsequently, a deep autoencoder neural network and a deep long short-term memory (LSTM) network were integrated to categories the ship-radiated noise spectrogram. Additionally, a multimodal deep learning approach was put forth, using the ship-radiated noise spectrogram as the input for the acoustic modality.

A combined model for underwater acoustic target detection was put forth, based on a waveform and a T-F mode. Two divisions of the model are T-F and wave branches. As the core of the T-F branch, ConvNeXt was an advanced lightweight deep neural network. To train the joint model, a synchronous deep mutual learning approach based on deep mutual learning (DML) was presented. In order to recognize underwater acoustic targets, the suggested joint model may automatically learn to extract deep characteristics from each branch and incorporate heterogeneous modes. Real-world scenario datasets were utilized to confirm the efficacy of the suggested model. The optimal recognition accuracy of 85.20 percent was attained by the joint model utilizing synchronous deep mutual learning. This was enhanced by 2.05 percent using the MSRDN and 1.25 percent using the separable convolution autoencoder (SCAE). Wavebranch's core was suggested to be a lightweight MSRDN. The primary focus was on designing a lightweight framework based on the original redundancy architecture. Two datasets were used to evaluate the performance of proposed model. The ONC dataset was created by Ocean Networks Canada. Following the label system, the ONC dataset was divided into four target groups, each with roughly 62.5 hours of recordings. The first twelve months' worth of recordings are utilized for training, while the rest are used for testing. A dataset for underwater acoustic benchmarking called DeepShip was been proposed. DeepShip was made up of 47 hours and 4 minutes of actual underwater recordings from 265 distinct ships that fall into four categories. Another source of the data was Ocean Networks Canada.

To improve the recognition impact, they retrieved many features for feature fusion as the network input. They employed three characteristics that were frequently used in music theory in addition to more classic ones like Mel-spectrogram and Mel-Frequency Cepstral Coefficients: chromatogram, spectral contrast, and tonnetz. The methodology was tested on Ship-Ears dataset [5].

Hu et al. [6] proposed Deep convolutional neural networks and ELM, a technique for feature extraction and identification of data on underwater noise. Deep convolutional neural networks are neural networks that have two or more convolutional and sampling layers throughout the whole network.

Lou et al. [7] proposed a target recognition approach based on integrated characteristics with automated coding and reconstruction in order to categories ship-radiated noise signals. A feature extractor based on auto-encoding was created in the proposed recognition approach. The feature extractor automatically encodes the combined data of the power spectrum and demodulation spectrum of ship radiated noise without supervision and extracts the deep data structure layer by layer to produce the signal feature vector using the restricted Boltzmann machine (RBM). To achieve target recognition, a Back Propagation (BP) neural network receives the extracted feature vector. The extended sample set was created using a method of data augmentation developed by RBM auto-encoder, which boosts the performance of the recognition system.

Doan et al. [8] suggests a productive method for classifying UA signals that makes use of a dense convolutional neural network (CNN) that was skillfully built to automatically learn representative features without the need for domain transformation and feature engineering expertise. In practical terms, the network architecture utilising the skip-connection technique permits the network to reuse all previous feature maps generated at multiscale representations. In order to identify 12 kinds of UA signal, proposed model examines a deep neural network with dense architecture called underwater acoustic target classification DenseNet (UATC-DenseNet). A fibre receiver on the ground retransmits data from optical to local area network (LAN) transmission. Following that, the signal data are kept in a multistore device that has three hard drive discs. A switch in the multistore enables the surveillance system to access and examine either the stored data set or the real-time signal.

The data set was splitted into 70 percent for training and 30 percent for testing at random. In the initial investigate, where the size of 1-D kernels configured in the convolutional layers varies in the set, the impact of the kernel size parameter on the overall accuracy is thoroughly examined. The model classifies more accurately the higher the kernel size. Many temporal correlations are then recorded at various feature representations and subsequently merged using the astutely used skip-connection strategy in UATCDeepNet. Expanding the network to three convolutional blocks results in a notable improvement in accuracy, with an approximate improvement of 1.90 percent of the mean. The suggested network significantly outperforms several well-known machine learning techniques in the final experiment, such as random forest, DT with ten compact classification trees and KNN. Tenfold cross validation was carried out using the 14 features that the MFCC approach had collected. By means of a passive sonar system's performance evaluation on our real-world data set, the suggested CNN-based classifier achieves a classification rate of up to 98.85 percent.

Xiao et al. [9] an attention module was used to look into the neural network's inner workings and a target recognition attention-based neural network (ABNN) was developed for the pressure spectrogram with multi-source interference. To interpret the classification concept of DNN, an attention-based neural network was presented for underwater audio target detection. An attention module was included in the ABNN architectures as a preliminary step before a conventional DNN made up of connected layers. The attention module mines useful information from the input spectrum using a trainable attention vector layer, a Gaussian layer, and a merge layer. It then outputs attention maps in real time to show the frequency zones of importance. The fully linked layers function as a classifier to complete particular jobs, like target identification and detection. The performance was assessed using the South China dataset. An experiment measuring ship-radiated noise was carried out in the South China Sea's shallow waters. The experimental area's seabed topography. With their main and auxiliary engines deactivated, two experimental ships, A and B, floated at either end point or travelled at a speed of three meters per second along tracks 1 and 2. Both of the lines have a comparatively level bathymetry. Eighty meters below the surface, a hydrophone was set up as a submersible buoy. The source was located between 1-4 km away from the hydrophone, with a water depth of 120-134 meters. The spectrograms show that over 17 interfering vessels were found in the experimental area by the radar and automatic identification system (AIS), which has an 11 km detection range. In this study, target detection and recognition from a pressure spectrogram are carried out utilizing ABNN in the face of multi-source interference. Two-second windows were chosen, zero-padded, and short-time Fourier converted for a portion of the near-field signal data recorded at 12 kHz. As a result, the signal data were imported into a spectrum dataset of 1254 frames. Just 245 frequencies between 10 to 100 Hz were selected as the input features for the DNN due to the importance of low-frequency characteristics of ships. Based on data collected during a sea experiment in September 2020, the ABNN demonstrated a steady concentration on the target ship's frequency domain feature and suppressed background noise and interference from other marine vessels. The model controls the sensitivity of frequency components by weighting the input features with a dense layer of length 248 that was activated by SoftMax. The focus on target characteristics and suppression of multi-source interference, the visualization of problematic features during target detection or recognition, the ability to resolve multiple targets using only single-target data, and the model's suitability for use as a dedicated feature extraction model are the characteristics that define ABNN. As the training loss steadily drops throughout DNN training, the ABNN progressively concentrates its learning on the features that are highly connected with the training objectives. The ABNN performs well in multi-target resolution and target detection and recognition. The class was defined as the highest value among these components. The test datasets showed 98.0 percent and 97.3 percent detection accuracy after twenty thousand epochs.

Jiang et al. [1] proposed the modified DCGAN model to supplement data for targets with small sample sizes. They focused on investigating on the classification of underwater targets using a deep learning algorithm. We have presented the modified DCGAN model to enrich the underwater target dataset by creating fake data with high quality and variety based on genuine target data, thereby addressing the issues of short sample size and imbalanced categories of underwater target data. For underwater target categorization, we have presented the S-ResNet model, which combines CNN with SqueezeNet, a popular kind of lightweight neural network. They discovered that our suggested model achieves high classification accuracy at a considerable reduction in model complexity. The primary goal was to suggest a modified DCGAN model to supplement underwater target data, which might enhance the training stability and quality for underwater targets with a small dataset. Additionally, an S-ResNet model was presented in order to achieve good classification accuracy at a large reduction in model complexity. Furthermore, field tests involving five distinct categories of submerged targets have been conducted to confirm the efficacy of suggested models.

They had created a new fire module as the constructive unit block for the S-ResNet classification model, which was inspired by the concept of the SqueezeNet fire module. Through convolutional kernel decomposition and compression ratio hyperparameter, the S-ResNet classification model can further enhance the performance of quantitative neural networks without adding more CNN parameters. Less parameters are a benefit of the S-ResNet classification model over the traditional convolutional neural network. In the suggested model, SoftMax was combined with the cross-entropy loss function. Measured data from lake and sea trials were used to demonstrate the efficacy of the proposed models. Three distinct locations in China were used for the experiments, Jiao Zhou Bay in Shandong Province, Yangjiahe Reservoir in Shaanxi Province, and Danjiangkou Reservoir in Henan Province. Five distinct target types are represented by the data gathered: a motorboat, two different kinds of ferries, a speedboat, and a frogman. Naturally, with more generated trained data, the frogman's classification accuracy improves dramatically from 83.6 percent to 94.8 percent. Furthermore, despite the fact that no further trained data are produced for these targets, the classification accuracy of other targets was either maintained or increased by 4.3 to 2.5 percent. However, because of the extremely limited energy resources and hassle of changing batteries in an underwater environment, even though the proposed model has greatly decreased model complexity, the complexity still needs to be further lowered for practical implementations.

The researchers [10] proposed feature extraction using the GRU-CAE collaborative deep learning network, template creation, and template matching are included as three stages. The sample set of underwater acoustic targets includes a restricted number of categories of targets, each of which has a specific label. In addition, the sample set of underwater acoustic targets includes categories of unknown labels, which can be either finite or infinite. The goal of underwater acoustic target open set identification was to identify specific subsets of underwater acoustic targets while rejecting recognition of subsets with uncertain labels.

There are three basic steps in the underwater acoustic target open set recognition process using the GRU-CAE cooperative deep learning network. In order to extract the deep collaboration features, the GRU-CAE cooperative deep learning network must first be built, and its parameters must then be optimized by training. Deep collaborative features are the name given to the deep features. The feature template was built in the second step. The feature template was computed as the mean of the deep collaborative features of the training set samples. Thirdly, open set identification using template correspondence. European template matching distance, or Euclidean distance, was calculated between the test set's deep cooperative features and the feature template. The category of the test set samples was then determined by choosing the optimal thresholds. The primary determinant of underwater acoustic target open set identification performance was the intra- and inter-class compactness and separability of deep collaborating characteristics. Consequently, in order to extract deep collaborative features with improved intraclass compactness and interclass separability, GRU and CAE networks are chosen to construct a collaborative network. The performance of the suggested method's recognition was tested using experimental data from 5 different types of underwater sound targets.

Three forms of ship radiated noise and marine environmental noise are present in the training set data and the closed set test set; four types of ship radiated noise and marine environmental noise are present in the open set test set. The open set test set now includes a class of data E that did not exist in the training set. Additionally, distinct people from the same class are included in both the training and test sets, and the marine background noise in each set was captured at a different time. Every kind of ship emitted noise, and these data were gathered from several different ships. Every ship's radiated noise sample lasts for one second. The training set contains every ship class of the closed set recognition test. The three different collaborative network types have closed set identification accuracy that was 2 percent to nine percent greater than that of the CNN or GRU network models.

Authors of [11] developed a ResNet-based underwater acoustic target recognition (UATR) technique. The classic time-frequency (T-F) analysis method struggles to simultaneously extract different signal characteristics, therefore the proposed method uses a multi-window spectral analysis (MWSA) method to overcome this problem. For the classifier's input, MWSA performs multiple window processing to create spectrograms with various T-F resolutions. The methodology was tested on Ship-Ears dataset [5]. This dataset comprises of 90 acoustic recordings from 11 different ship types over a 15- to 10-minute period. They can be divided into four categories based on the types of ships, namely A, B, C, and D, and E for ambient noise, according to the annotation in the original dataset. A conditional deep convolutional generative adversarial network model was created for high-quality data augmentation due to the insufficient amount of ship-radiated noise sample.

In [13] proposed cross entropy loss function based on trigonometric function to solve imbalanced dataset. Authors in [14] proposed an underwater acoustic target multi-attribute correlation perception method based on deep learning. The attributes include ship types, ship size, propeller type, etc. Researchers [15] also focuses multi-attribute and using its power spectral density for signal energetics, temporal coherence for machinery tonal sound, and spectral coherence for propeller amplitude-modulated cavitation noise, the underwater sound emitted by a ship with a variable pitch propeller was examined and measured. Tonal signals that are frequency modulated are likewise described in terms of frequency variations.

Ayvaz et al. [12] proposed the zero-crossing extraction and used energy level detection in order to identify areas that were voiced or unvoiced in the recorded speech signal. The voiced signals were discovered and used for segmentation. Furthermore, each segmented window was subjected to the MFCC technique. The MFCC data that had been retrieved were also utilised as training inputs for neural networks. The Mel Frequency Cepstral Coefficients (MFCCs), which are employed in a number of speech processing approaches, could potentially serve as the foundation for the feature extraction method. For identifying underwater audio data, MFCCs may be modified [16].

Researchers in [22] suggested an approach based on CNN and ELM for the feature extraction and identification of underwater noise data. Depth convolution network was used to provide an autonomous underwater acoustic signal feature extraction approach. An extreme learning machine was the foundation of an underwater target recognition classifier. Convolution neural networks are capable of performing both feature extraction and classification, but their primary function depends on a full connection layer that was trained using gradient descent; as a result, their ability to generalize is limited and suboptimal, necessitating the use of an extreme learning machine (ELM) during the classification stage. CNN removes the fully connected layers after first learning robust and deep features. To perform an excellent classification, ten ELM fed with CNN features was utilized as the classifier. In comparison to the conventional Mel frequency central coefficients and Hilbert-Huang feature, the recognition rate improved significantly in experiments conducted on the real data set of civil ships, yielding a 93.04 percent recognition rate. Xinwei Luo et al. [23] gives the comprehensive survey of underwater acoustic target classification methods. The current tendency in UATR development was to mix machine learning techniques with manual characteristics, owing to the scarcity of training data. In the study, feature extraction techniques and their corresponding properties for underwater audio target recognition were presented.

Sometimes underwater dataset is limited and need to expand for the features extraction process and proper evaluation. Ashraf H et al. [24] focuses on expanding dataset using generative adversarial network and then extract features. Researchers also used basic audio features to extract useful features for the classification of audio files [18-21].

2.2.2 Siamese Network

A popular method of neural network architecture that is intended to recognize and distinguish between input pairs is the Siamese network. It is frequently employed for a variety of applications, including facial recognition, signature verification, picture similarity, classification and identification. The network as shown in figure 2.1 is made up of two identical subnetworks called twins that have the same characteristics and weights. The two input samples (anchor and positive) are processed by these subnetworks, which result in embeddings. Making the embeddings of dissimilar inputs (negative pairs) far apart and comparable inputs (positive pairs) close to one other is the aim [17].

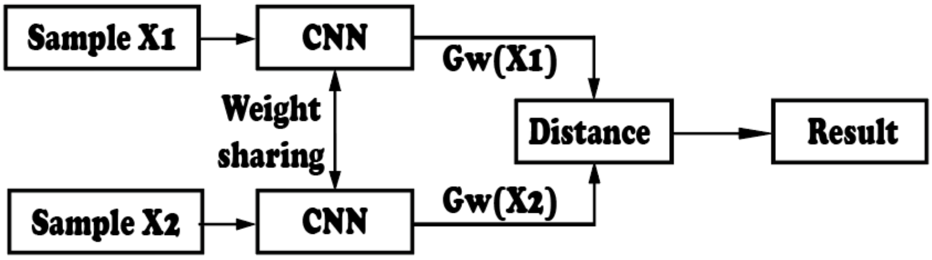

Figure 2.1: Siamese Network architecture [17]

An expansion of the Siamese network, the Siamese triplet network is made to learn embeddings for three input samples: an anchor, a positive example that resembles the anchor, and a negative example that differs from the anchor. The objective was to maximize the distance in the embedding space between the anchor and the negative example and minimize the distance between the anchor and the positive example. Common applications of this kind of network include picture retrieval, face recognition, and human re-identification.

Dali Liu et al. [17] proposed Siamese Network that includes two identical one-dimensional convolutional neural networks, which was capable of identifying envelope modulation on noise (DEMON) spectra of noise emitted by underwater targets. While the conditions of the samples that were obtained were quite uniform, the parameters of the underwater samples varied greatly. Conventional underwater target recognition involves expensive multi-state sample network training. This article used samples from a single state to train the network. Being able to recognize samples with various parameters was anticipated. Target datasets with various Doppler shifts, signal-to-noise ratios, and interference levels were created in order to assess how well the suggested Siamese network generalized. The experimental results demonstrated that the suggested network's classification accuracy reached 95.3 percent when it came to identifying samples with Doppler shifts. The classification accuracy for SNRs was 85.5 percent. The suggested model's exceptional generalizability demonstrates its applicability for real-world engineering applications.

Xingping Dong et al. [25] proposed a unique triplet loss to extract expressive deep feature for object tracking by incorporating it into the Siamese network architecture in place of pairwise loss for training. Their method was able to combine the original samples with additional parts for training, resulting in a more robust feature without the need for additional inputs. In addition, they provide a theoretical study that combines back-propagation with gradient comparison to demonstrate the efficacy of our approach. We use the suggested triple loss for three Siamese network-based real-time trackers in our studies. Furthermore, our variations outperform baseline trackers on a number of well-known tracking criteria, operating at almost the same frame rate and achieving comparable accuracy to more modern, state-of-the-art real-time.

A new approach was put up by Bhrugu Bhatt et al. [26] to use Few-Shot Learning, a machine learning technique that performs binary classification using a small support set, to tackle a two-fold challenge. Due to patient privacy concerns, traditional machine learning techniques use very large medical training datasets, which can result in expensive computing costs. In addition, convolutional neural networks (CNNs) like VGG19 and GoogLeNet were contrasted to see if FSL on the CNMC dataset resulted in any appreciable improvement. After 10 epochs of training, our Siamese network currently has an 85 percent testing accuracy. But transfer learning has also demonstrated that it was possible to achieve excellent accuracy rates with little data and less computationally demanding training. GoogLeNet and VGG-19, which were trained for 50 epochs with image transformation preprocessing, yielded accuracy rates of 99 percent and 96 percent, respectively.

Siamese network achieved about 99 percent accuracy in different tasks such as face recognition, detection of remote sensing images, verification system [27-31]. The goal of auditory-to-score alignment was to produce an exact mapping between a performance's audio and the composition's score. Conventional alignment techniques rely on Dynamic Time Warping (DTW) and use manually designed characteristics that aren't adjustable to suit various acoustic circumstances. Ruchit Agrawal et al. [32] proposed the technique that uses audio to assess alignment and learnt frame similarity to get over this restriction. Their main focus was the offline audio to piano score alignment. Experiments conducted on music data from various acoustic settings show that our method provides stable alignments while being flexible enough to be used to different domains. It also achieves higher alignment accuracy than a traditional DTW-based method that relies on handcrafted characteristics.

Abbas A et al. [33] presented a method for kinship recognition that combines a Siamese network with a pre-trained LAT model. The comprehensive experimental outcomes were employed to verify the efficacy of their suggested paradigm. Additionally, their model surpassed the state-of-the-art models that were previously used using a similar technique, as seen by the comparative study with previously conducted studies, yielding an overall accuracy of 76.38 percent. Loris Nanni et al. [34] create a dissimilarity space by combining a Siamese neural network with several clustering methods. This space was then utilized to train an SVM for automated animal audio classification. The two animal audio datasets that were utilized are the publicly available bird and cat noises. Utilizing several clustering techniques, we reduce the dataset's spectrograms to a number of centroids, which are then utilized to create the dissimilarity space via the Siamese network. Their research demonstrates that the suggested dissimilarity space-based strategy works effectively on both classification tasks without the need for ad hoc clustering method optimization.

In order to enhance the fine-tuning performance, a semi-supervised strategy based on the similarity of deep features was proposed by Shengzhao Tian et al. [35] for mining and labelling partial unlabeled samples. Four underwater acoustic target recognition models have their performance baselines created based on a series of short sample datasets with varying amounts of labelled data. Using the suggested framework significantly enhances the recognition effect of four models as compared to the baselines. Xiao Cheng et al. [36] proposed an innovative deep learning technique that abstracts time-domain signal characteristics was examined using a hybrid routing network. Because the learned properties of various branches are exchanged, the used network, which has several routing structures and possibilities for the auxiliary branch, promotes outstanding effects. The experiment demonstrates that the employed network has greater advantages when it comes to the task of classifying underwater signals. Honka T et al. [37] proposed a method on one-shot learning with Siamese networks for audio in the environment and results proved that it was the valid approach for classifying the environment audio. Siamese network achieves 98 percent in underwater visual loop detection [38]. In order to address the issues with the current conventional MOTs, Lee et al. [39] suggest a new MOT system. To overcome the structural simplicity, the feature pyramid

Siamese network was suggested. The feature pyramid network (FPN) served as the model for the FPSN, which expands the Siamese network by creating a new multi-level discriminative feature and applying FPN to the basic Siamese design. As a result, FPSN-MOT takes motion information into account in addition to appearance features. Lastly, the public MOT challenge benchmark problems are used to test FPSN MOT, and the results are compared to those of other techniques.

Their study suggests a Siamese distillation network to increase tracker efficacy in maritime conditions. Comprehensive experimental findings show that this network works better than other trackers in terms of accuracy. To be more precise, this network outperformed other Siamese networks with an accuracy value of 0.612, which increased its performance over the baseline network by 2.5 percent [40]. Dawkins et al. [41] introduces the FishTrack23 dataset, which offers a vast amount of expertly annotated fish ground truth tracks in images and videos that were gathered from a variety of diverse backgrounds, places, gathering circumstances, and organizations. A lightweight Siamese network based on a hybrid excitation model was suggested by [42] as a solution to the mismatch between deep learning models and limited characteristics underwater objects. Yin et al. [43] uses five categories to group 17 well-known trackers, and five data sets to evaluate each tracker's performance using Siamese Network. Tracking algorithm was used for the Siamese Region Proposal Network [44].

Zhang et al. [45] proposed a new Siamese anchor free network that utilizes an enhanced head network and crisscross attention. The template image and search region's features are extracted using ResNet-50, and the feature maps were then fed into a recurrent attention module to improve discrimination. Their modified head network receives the improved feature maps as input. Regression branches were used to remove low quality bounding boxes. Chen et al. [46] proposed pseudo-Siamese neural network that was specialized for underwater noise characteristics and was based on self-supervised picture denoising. In addition, the pyramid hierarchical structure was intended to simultaneously take into account feature information under various scales, which helps to reduce image noise. Their model was capable of greatly reducing the noise resulting from the scattering of undersea contaminants. [47-52] Siamese Network shows state-of-art performance in object tracking.

The authors of [53] demonstrates that employing a variation of the triplet loss to achieve end-to-end deep metric learning works far better than the majority of other published approaches for both pretrained and newly trained models. Three primary sections comprise their experimental evaluation. The initial segment assesses various iterations of the triple loss, encompassing certain hyper-parameters, and determines the optimal configuration for individual ReID. They conduct the evaluation on a train/validation split that they construct using the MARS training set. The performance they can achieve depending on the chosen triplet loss variant. On the CUHK03, Market-1501, and MARS test sets, they presented state-of-the-art results using both a pretrained and a freshly trained network. The many triple training variations are tested in their early experiments. They randomly sampled a validation set of roughly twenty percent of people from the MARS training set so as to avoid doing model-selection on the test set, leaving the remaining eighty percent people for training. The researcher conducted all of these tests using the smaller LuNet, which was trained from scratch on photos that had been downscaled by a factor of two in order to make this exploration tractable. In their trials, they did not undertake any data augmentation because their aim was to investigate triplet loss formulations rather than achieve optimal performance.

One of the greatest text classification techniques was the kNN algorithm, which was a well-known pattern recognition technique. It was among the most straightforward machine learning methods for categorization. The paper [54] provides an overview of the kNN algorithm and its related literature, delves into the algorithm's concept, steps, and implementation code, and evaluates the benefits and drawbacks of the several improvement strategies. The evolution of the kNN algorithm, as well as significant published publications, were also introduced. Main concept of kNN algorithm was, the k neighbors who were closest to the sample data to be sorted were identified by calculating the distance between the sample to be sorted and the training sample of the known category. The categories to which the neighbors belong decide which categories the sample data to be sorted falls into. The selection of the value of k was crucial.

If the value of k was too large and the sample's classification belongs to the training set, which contains fewer data classes, then the fact was not similar to the training set when choosing k neighbors. Similarly, if the K value selection was too small, the number of neighbors was too small, which would reduce the classification accuracy but also amplify the noise data interference. The algorithm was best for classification tasks as well as regression.

Unlike the image recognition field, underwater audio target detection lacks the vast quantity of high-quality labelled samples needed to train robust deep neural networks. Additionally, gathering and annotating a large number of base class data in advance was challenging. Consequently, it was challenging to utilize traditional few-shot learning techniques for underwater audio target detection. The research in [55] proposed learning framework for underwater acoustic target recognition model with few data, modelled after advanced supervised learning frameworks. In the meantime, a semi-supervised adjusting approach based on the similarity of deep features was proposed to enhance the performance by mining and labelling of partly unlabeled samples. Four underwater acoustic target recognition models had their performance baselines created based on a series of short sample datasets with varying amounts of labelled data. Using the suggested framework significantly enhances the recognition effect of four models as compared to the baselines. In particular, the joint model's recognition accuracy was improved from the baselines by about two percent to twelve percent.

Model dependence on the number of labelled samples can be effectively reduced when the model performs better on just ten percent of the labelled data than it does on the entire dataset. The issue of underwater acoustic target recognition lacking labelled samples was resolved.

2.3 Applications of Siamese Network

Siamese network-based underwater ship-engine classification has great potential for a wide range of applications across multiple sectors. A key use case is maritime security, where precise categorization of ship-engine noises can help identify and track questionable vessel operations, unapproved incursions, or possible security risks in delicate maritime areas or territorial seas. Siamese network-based categorization systems help to improve maritime domain awareness and protect vital maritime assets and infrastructure by offering early detection and alerting capabilities.

Underwater ship-engine categorization also has uses in environmental protection and monitoring. Researchers and environmental authorities can evaluate the effects of maritime traffic on marine ecosystems, identify regions of high ecological importance, and put policies in place to reduce pollution, habitat degradation, and disturbance of marine species by examining the acoustic signatures produced by ship engines. Furthermore, by categorizing ship-engine noises, environmental standards can be more easily monitored and enforced, promoting environmentally friendly maritime practices and reducing the negative impacts of human activity on marine habitats.

Siamese network-based classification algorithms are also useful for scientific research, monitoring fisheries, and managing marine resources. Authorities can effectively detect and control maritime traffic, enforce fishing rules, and evaluate the effects of anthropogenic activities on fish stocks and marine environments by accurately distinguishing ship types and activities based on their acoustic signatures. Furthermore, underwater ship-engine classification advances marine science, oceanography, and underwater technology by offering insightful information about marine spatial patterns, ocean currents, and underwater acoustic environments. All things considered, the great range of applications highlights the significance and possible social advantages of underwater ship-engine classification with Siamese networks.

2.4 Comparison of Underwater Classification Techniques

Researchers have conducted much study in the field of underwater classification methods and classifiers, as evidenced by the survey and literature mentioned above. Feature extraction techniques have greater impact on model’s performance. However, the selection of dataset and classifier have also impact on accuracy. Table below shows the summary of proposed models, dataset selection and their limitations.

Table 2.1: Summary of models

Ref	Proposed Model	Dataset	Limitation
[3]	Feature extraction mel-sepctogram+ 1st derivative+ 2nd derivate Classification by CNN CNN acts as classifier	ShipEars data set including 11 types of ships, 5 classes This dataset comprises of 90 acoustic recordings from 11 different ship types over a 15- to 10-minute period. They can be divided into four categories based on the types of ships, namely A, B, C, and D, and E for ambient noise, according to the annotation in the original dataset.	When using the first and second derivatives of the Mel-spectrogram to integrate it for feature extraction, dimensionality and computational complexity are increased, which may result in overfitting. Because derivatives are susceptible to noise, interpretability and classification performance may suffer. This method may lose information and ignore fine-grained details while capturing temporal dynamics, which could reduce the ability of derived features to discriminate.
[2]	Five Feature extraction techniques Classification using Joint CNN-LSTM		The intricate architecture and large number of parameters in the CNN-LSTM combined model for classification may cause overfitting. Its scalability and applicability in real-time applications may however be limited by the substantial computer resources needed for training and inference.
[4]	Log Mel+ MFCC +CCTZ as feature extraction method ResNet as classifier		ResNet-18 because of its complex architecture, it needs a lot of memory and processing power.
[5]	Cepstral coefficients as feature extraction method GMM as classifier		The efficacy of the GMM classifier is limited for non-Gaussian or complicated data distributions since it assumes a Gaussian distribution within each class.
[7]	Restricted Boltzmann Machine as feature extraction		RBMs' computationally demanding training procedure and sensitivity to hyperparameters might cause them to converge slowly and have trouble training with large-scale datasets.

[11]	ResNet based UTAR Multi-window spectral analysis as feature extraction		Overfitting may result from multi-window spectral analysis's increased feature dimensionality and computational overhead. Furthermore, it can be difficult to choose the right window sizes and overlaps, which affects how robust the extracted features are across various datasets.
[6]	CNN ad ELM as feature extraction technique	Civil ship dataset	The random initialization of hidden layer weights is a crucial component of ELM's performance, and it can result in less-than-ideal outcomes or uneven performance between runs.
[1]	DCGAN model S-ResNet as classifier	There were five distinct target types: a motorboat, a speedboat, two types of ferries, and a frogman.	Because there are extremely little energy supplies available and changing batteries in an underwater environment is inconvenient, the complexity was not fully decreased for practical applications.
[8]	UATC-DenseNet model	11 UA signals, one with noisy blank signal, consisting 4096 samples	The dense connectivity network of the UATC-DenseNet model may lead to greater computational complexity and memory needs, which could limit its scalability to big datasets or resource-constrained contexts.
[9]	Attention based neural network	Ship A and Ship B audio recorded in South China sea, consisting 1254 frames	Due to the added attention mechanisms, attention-based neural networks may be more computationally expensive and difficult to train, which could restrict their scalability and efficiency.
[10]	GRU-CAE collaborative deep learning network	5 kinds of acoustic targets	The GRU-CAE collaborative deep learning network may experience difficulties optimizing the recurrent and convolutional components simultaneously, which could result in inferior performance or possible convergence problems.

[26]	Few-shot learning, VGG-19	CNMC (cancer patients' dataset)	VGG-19's deep architecture and numerous parameters may result in high computational costs and memory needs, making it impractical for real-time or resource-constrained applications.
[34]	A Siamese neural network + various clustering methods SVM as classifier	Bird and cat audios	Have difficulties interpreting the learnt representations and figuring out the ideal number of clusters, which could result in less interpretable models and worse-than-ideal clustering outcomes. Furthermore, adding computational overhead and complexity by combining Siamese networks with clustering may prevent scalability for big datasets or real-time applications.

2.5 Research Gaps and Directions

In the area of processing underwater acoustic signals, the classification methodology of these signals has always been an important study topic. Although comprehensive research has been made in this field, there are still some research gaps and directions that can be investigate. Through the extensive review on literature, here are some:

i) **Dataset Challenges:** Addressing the scarcity of extensive and varied labelled datasets for the classification of underwater acoustic classification. Usually, the dataset is not available for security reasons. Establishing and disseminating uniform datasets for the purpose of comparing various techniques and methods.

ii) **Adaptation to Underwater Environments:** Building deep learning models with features tailored to address the difficulties posed by underwater acoustic data, including signal attenuation, fluctuating environmental conditions, and background noise

iii) Transfer Learning: Examining how well pre-trained models perform on generic acoustic tasks when using transfer learning approaches to improve underwater acoustic categorization performance.

iv) Robust Feature Extraction: Enhancing the flexibility of feature extraction techniques to shifts in source-receiver geometry, species behavior, and environmental fluctuations are key components of robust feature extraction. Diversify enough the features that can be easily classified by classifier. Establishing new methods for feature extraction for better performance.

v) User Feedback Integration: Creating systems that are capable of taking into account user input and using human knowledge to enhance model performance and adjust to shifting underwater circumstances.

2.6 Summary

The latest techniques for underwater classification and siamese network for recognition is discussed in this chapter. We have done the critical analysis of various techniques and conclude that the research mainly emphasis on modification of feature extraction techniques. Integration of different techniques lead to higher accuracy, however, increase in cost, time and complexity. In addition, a tabular representation of the models, datasets used, and drawbacks has been provided for comparison. Research gaps in earlier studies have also been mentioned, along with topics for further research.

CHAPTER 3

METHODOLOGY

3.1 Overview

A new method of classification of ship-engine audios based on Siamese Network is designed and developed in this chapter. The primary objective of this study is to develop a technique for diversifying the features using a simple network so that the classifier can easily classify the correct labels for audios. Reducing the concept of multiple feature extraction techniques and deploying a network which will be used as feature extraction. The performance evaluation is checked on ShipEars [5] dataset. The research strategy outlined in this chapter is thought of as an initial step that provides guidance for subsequent actions in order to meet predetermined goals.

3.2 Proposed USCSN Model

The proposed underwater ship-engine classification using siamese network has been used take ship-radiated noise as input and convert them into spectrograms. The short time fourier transform (STFT), which produces a spectrogram, a two-dimensional (2D) image with the frequency content of the noise signal changing over time, will create the feature space [3]. STFT (Short-Time Fourier Transform) is more suitable for spectrograms as it is the compromise between time and frequency resolution that works with non-stationary signals. The STFT affords a so-called time-frequency localization, which works for high frequencies with better time resolution and low frequencies with better frequency resolution. While the Wigner-Ville Distribution (WVD) can get a high-resolution spectrogram but may involve cross-term interferences, STFT is free of this cross-terms. On the other hand, the MFCCs [2] are indeed effective tool for speech recognition, but they are less straightforward for visualizing frequency content over time as opposed to STFT which delivers the job very well. The spectrograms and audio chunks have been sent to feature extraction module.

Siamese network has been acting as feature extraction and [17] extract features and give results in form of 1-D vector. The features extracted by the Siamese network from spectrograms of ship engine sounds can represent various aspects of the audio signals that are relevant for distinguishing different engine sounds. These features may capture temporal patterns such as the rhythm, repetition, or duration of certain engine sounds. For example, different engine types might have distinct patterns in how their components interact over time. These features could represent frequency characteristics such as the dominant frequencies, harmonics, or spectral shapes associated with different engine types. Certain engine types might produce characteristic frequency peaks or distributions due to their machinery design and operation.

Additionally, Features might encode information about the amplitude variations over time or frequency, which could reflect differences in engine behavior, power output, or environmental conditions. The features from siamese network may also capture transient events or unique spectral signatures that occur during specific engine operations, such as starting, accelerating, decelerating, or idling. Depending on the network architecture and training data, features might also indirectly encode non-acoustic factors such as engine size, load, condition, or surrounding environment, if these factors correlate with the observed acoustic patterns. Interpreting these features by various classification techniques on the basis of similarities involve analyzing their distributions, correlations, and relationships between them. The correct labels for each class will be easily classified by the classifier, and the results will be displayed as shown in figure 3.1. The proposed model has been applied on ShipEars dataset to evaluate the performance.

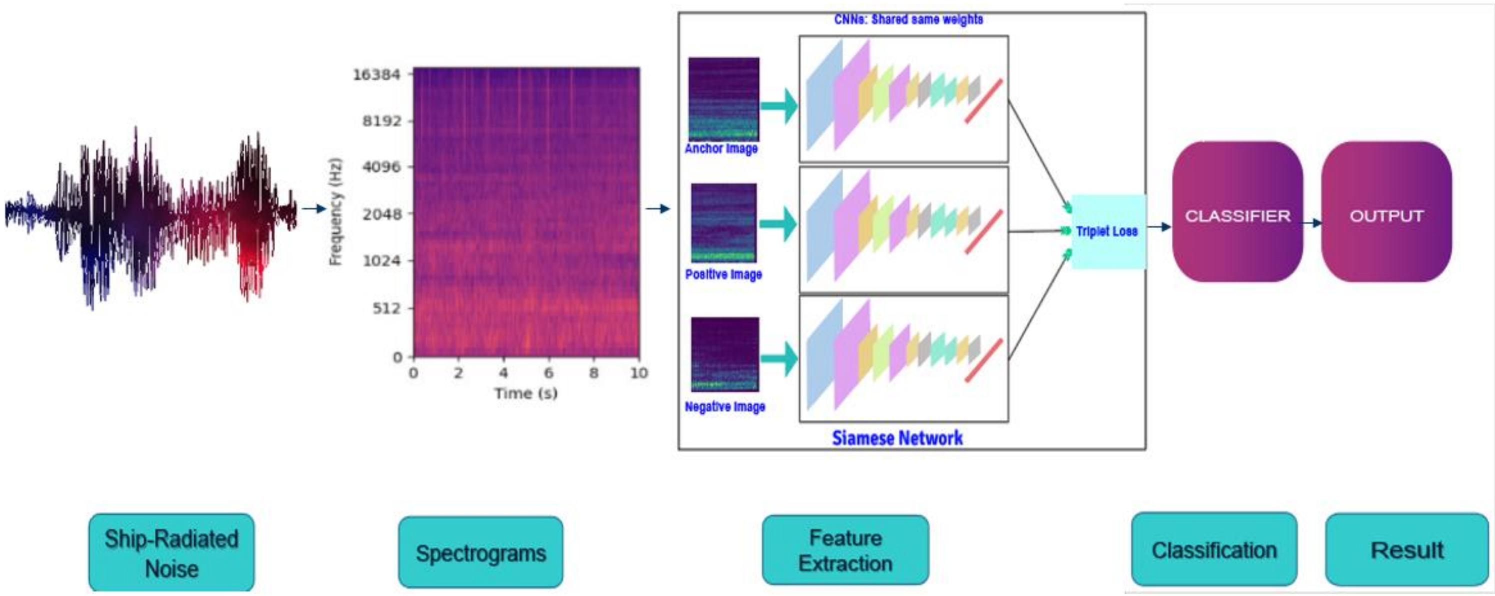

Figure 3.1: Proposed USCUS model

3.3 Dataset Description

Audio recordings of ShipEars dataset [5] were taken at several locations along the Spanish Atlantic coast in northwest Spain between the autumn of 2012 and the summer of 2013. A wide variety of vessels from the docks, such as tugboats, pilot boats, yachts, small sailboats, fishing boats, ocean liners, ferries of varying sizes, containers, ro-ros, total of 90 recordings. Recordings from hydrophones positioned from docks to record various vessel speed noises and cavitation noises associated with docking or undocking movements were added to ShipEar. High background noise that is often audible is caused by waves pounding on the port infrastructure. Additionally, recordings of ships operating normally were added to ShipsEar. The hydrophones were deployed as shown in figure 3.2 by an auxiliary vessel, and the recordings were scheduled based on vessel movement data that was gathered from the port authorities and the automatic identification system for vessels.

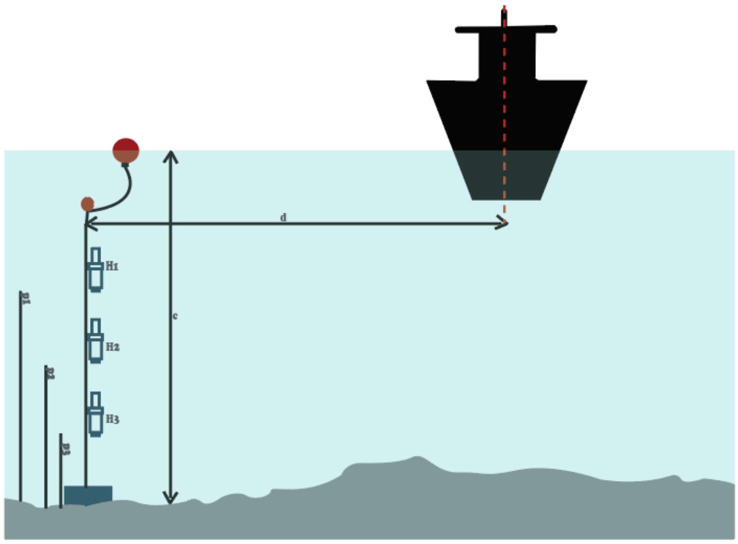

Figure 3.2: Hydrophone setup for underwater recordings of vessel noise [5]

ShipEars real-time dataset was available on internet. Other datasets were expensive and some were not publicly available. We have modified the ShipEars [5] dataset. We have divided the 90 recordings into 12 different classes instead of 5 classes in original ShipEars dataset; tugboat, dredger, mussel boat, trawler, motorboat, yacht, pilot boat, sailboat, passenger boat, ocean-Liner, ro-ro, ambient noise as shown in table 3.1.

Table 3.1: Underwater ships categorization

Classes	**Ships**
Class 1	Tugboat
Class 2	Dredger
Class 3	Mussel boat
Class 4	Trawler
Class 5	Motorboat
Class 6	Yacht
Class 7	Pilot boat
Class 8	Sailboat
Class 9	Passenger boat
Class 10	Ocean-Liner

Class 11	Ro-ro
Class 12	Ambient noise

3.4 Operational Framework

Siamese network used in this study has been used as feature extraction technique. Steps starting by analysis phase, design and development phase, and performance evaluation as shown in Figure 3.3 is the operational framework of this study. First, in the analysis stage, dataset of ship-engine audios is divided into 2 sec chunks. The spectrograms are generated using STFT. Secondly, prepare dataset for implementation. Generate triplets (three pairs) of dataset (anchor, positive example of a class, negative example of class). Split dataset randomly for training and testing. Then, implement the Siamese Network to extract the features and send them for classification to the classifier. In last phase of framework; performance evaluation of proposed model by using SNE-plots, confusion matrix, ROC curve, accuracy, precision, recall, f1-score.

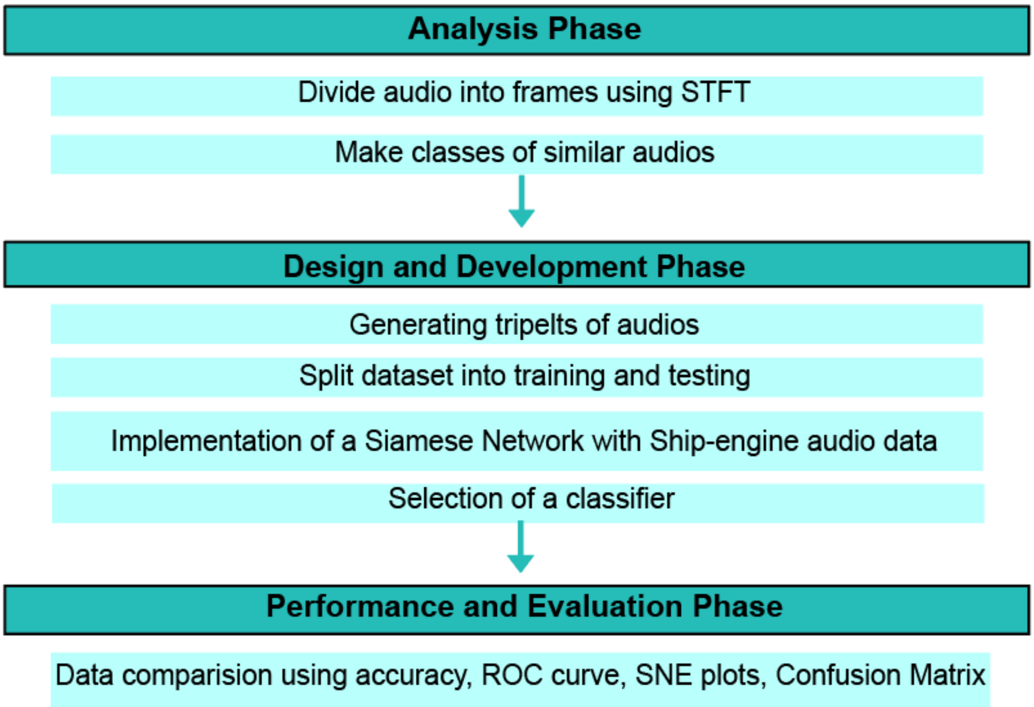

Figure 3.3: Operational Framework of the Research

Dataset contains 1 to 2 minutes of audios which are difficult to use and analyze for system. First the audios are divided to 2 seconds chunks and spectrograms are generated. A spectrogram is a two-dimensional representation of a signal in which the intensity denotes the amplitude or power of the frequencies contained in the signal, the x-axis indicates time, and the y-axis represents frequency as shown in figure 3.4. Second step of this phase is to make classes of spectrograms of audios and normalize the dataset using mean and variance.

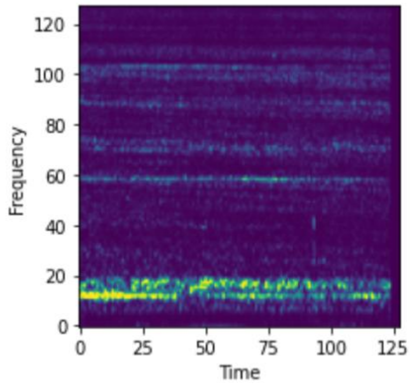

Figure 3.4 Spectrogram

3.5 Research Design and Development Phase

The design and development of proposed model is divided into steps; generating triplets, split dataset into training and testing, implement siamese network, send data to classifier.

3.5.1 Generating Triplets

To prepare dataset of ShipEars [5] for implementation, first triplets are generated. Triplets are basically the 3 pairs of spectrograms. Triplets in a Siamese network are created by choosing positive, negative, and anchor samples from the dataset. Anchor image is randomly selected image from dataset. Positive image is the true example of the class. Negative image is the false or negative example of the class. There are three examples of triplets listed in figure 3.5. 1st triplet is example of class 1, similarly 2nd triplet is example of class 2 and is 3rd triplet is example of class 3. The training objective for our model is to reduce the distance between the anchor and the positive sample and increase the distance between the anchor and the negative sample. Through this iterative process, the network will create embeddings that better reflect the relationships between similar cases in the input, which improves its performance in tasks like similarity retrieval and classification.

Because of the particular difficulties associated with the underwater environment, handling missing values in underwater datasets which frequently comprise sensor data or environmental measurements requires careful thought. One method is to estimate missing values by applying domain-specific imputation techniques, such as utilising established physical or environmental correlations. When working with time-series data, temporal interpolation can be helpful since it allows missing values to be inferred using the patterns of nearby time points. The audio data was extracted from continuous audio signals recorded from hydrophones. In cases where the audio data is missing (e.g., a few bits are missing), two strategies were followed. If the data is missing for longer durations, that portion of the audio was skipped during segmentation into chunks. For cases where the audio data is missing for less than one second, time interpolation methods were used.

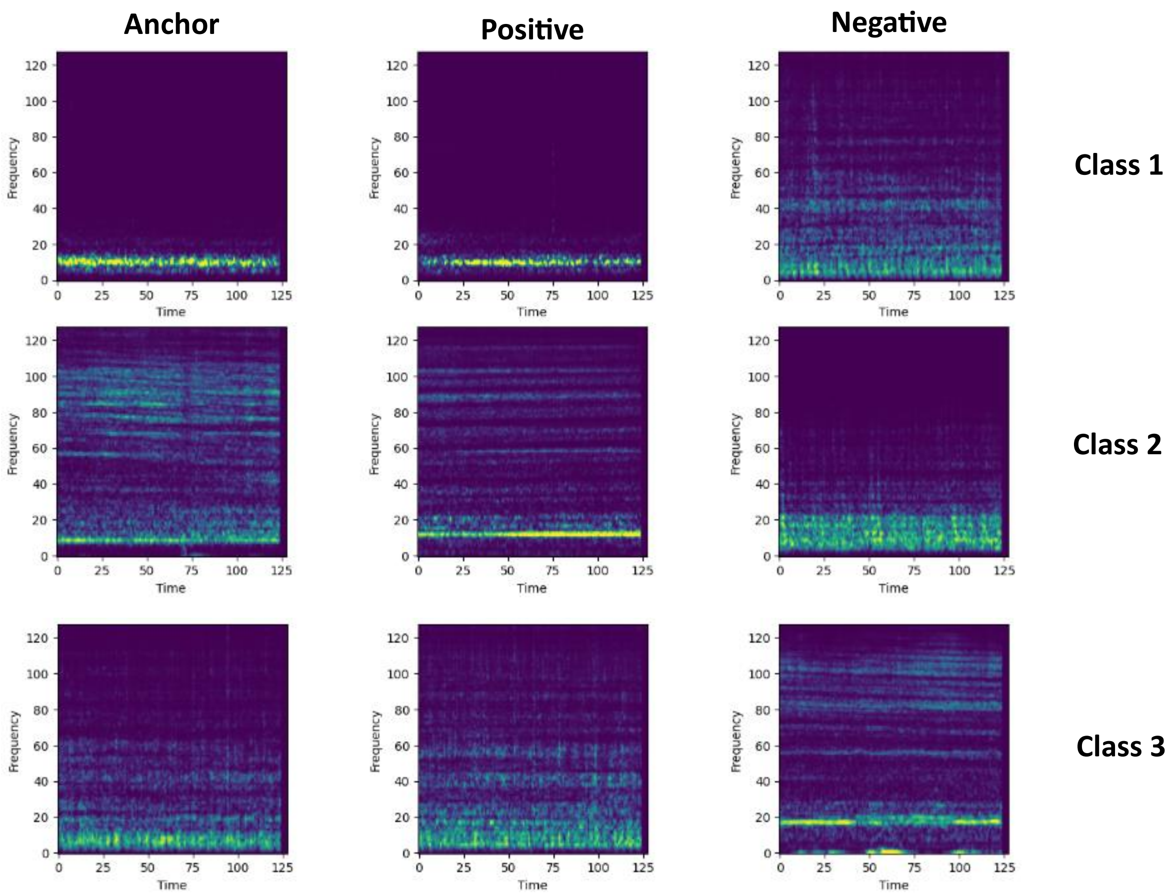

Figure 3.5 Visualization of Triplets

3.5.2 Split Dataset into Training and Testing

After generating the spectrograms, dataset then split into training and testing for future work. Eighty percent of data is randomly divided into training and remaining twenty percent of data is divided into testing.

3.5.3 Implementation

All three images have been passed through the same conventional neural network (CNN) in siamese network and shares same weights. The proposed structure of CNN network is shown in table 3.2.

Table 3.2: Layer detail of Siamese Network

Layer (type)	Output Shape	Parameters
Conv2D	(None,63,63,32)	320
MaxPooling2D	(None,31,31,32)	0
Dropout	(None,31,31,32)	0
Conv2D	(None,15,15,64)	18496
MaxPooling2D	(None,7,7,64)	0
Dropout	(None,7,7,64)	0
Flatten	(None,3136)	0
Dense	(None,1024)	3212288

Following by 3 columns, layer, output shape and parameters. The layers column shows layers used in our model and its type. Whereas Output shape column and its parameters explains the batch size, spatial dimensions and number of filters or channels produced by the convolution. However, parameter column shows the parameters used in each layer. First, the convolution layer (Conv2D) with kernel size of 3x3, strides 2x2 and activation function Rectified Linear Unit **(**ReLu) is applied to input image. Output shape of Conv2D

(None,63,63,32) Conv2D, where "None" indicates the batch size, 63x63 are the spatial dimensions of feature maps and 32 represents the number of filters or channels. Next is max pooling layer (MaxPooling2D) with pool size of 2x2. This layer uses max pooling to decrease the feature maps' spatial dimensions. (None, 31, 31, 32) is the output shape that keeps the number of channels constant but reduces the spatial dimensions in half. Then Dropout of 0.25, is applied to randomly set a portion of input units to zero in order to prevent overfitting. However, no modification in the proportions of output shape. Conv2D of 64 with kernel size of 3x3, strides 2x2 and activation function Rectified Linear Unit (ReLu). Yet another convolutional layer with an increased filter. (None, 15, 15, 64) increases channel depth while decreasing spatial dimensions even more. Then MaxPooling2D with pool size of 2x2 decreases spatial dimensions even further to 7x7 while keeping channel depth. Again, dropout of 0.25 for regularization. Flatten layer that creates a 1D vector from the 3D output that may be fed into a fully connected layer. Output: (None, 3136) gives the 7x7x64 tensor a length of 3136 by flattening it. Then the dense layer which is the fully connected layer of 1024 neurons with activation function Rectified Linear Unit (ReLu). Dense layer prepares the vector for future processing by transforming it into a 1024-dimensional higher-dimensional space after it has been flattened. Moreover, total parameters used in siamese network are 3231104 in which all the parameters are trainable and non-trainable parameters are zero.

The input image's convolutional layers extract progressively more abstract features, while the max pooling layers down sample the spatial dimensions to concentrate on the most significant features. In order to avoid overfitting, dropout layers add regularization. Ultimately, a dense layer is applied to the flattened vector to produce a high-dimensional embedding, which is where the network learns to identify similarities between images. Each layer's output shape, which is essential for comprehending the network's computational flow and design, represents the tensor dimensions following the application of the associated operations.

Triplet loss function has been applied after all this process. Triplet loss is a vital learning component in Siamese networks, helping the network discover embeddings that accurately represent the similarity relationships between input samples. The network is motivated to minimize the intra-class distance while maximizing the inter-class distance in the learned embedding space by use of the triplet loss function, which compares the distances between anchor, positive, and negative examples. In order to facilitate tasks like classification or similarity retrieval, this optimization process pushes the network to map similar objects closer together and dissimilar items farther apart. As a result, the triplet loss guides the network in identifying and representing significant links among input data, which improves the network's performance in a variety of similarity-based tasks. The triplet loss function is written in equation 3.2 [63]. In this triplet loss function for a Siamese network, d(a,p) denotes the distance between the anchor "a" and positive "p" sample, d(a,n) the distance between the anchor and negative "n" sample and d(a,n_{hard}) is the distance between the anchor and a hard negative n_{hard} sample. The margin ensures that the distance between positive pairs is larger than the distance of negative pairs by some buffer space. This loss penalizes the network if d(a,p) is not relatively smaller than d(a,n) and d(a, n_{hard}) is not relatively larger than d(a,p). The goal is to maintain the distance above the margin between the anchor and the positive sample while maximizing the distance between the anchor and the hardest negative sample [53].

$$max(d(a,p) - d(a,n) + margin, 0) + max(d(a, n\_hard) - d(a,p) + margin, 0) \tag{3.2}$$

Another CNN model been implemented without the Siamese network to extract useful features in image classification. We can easily compare the performance of two models, first with siamese network, then CNN network without siamese network using the accuracy. Loss function used in this CNN model is categorical cross entropy. The categorical cross-entropy loss function is used to measure the model's performance through which the model's predictions are compared with the actual class labels and used to optimize the model. This loss function is suitable for multi-class classification problems since it enables the model make necessary changes to its weights for more accurate prediction. Categorical cross-entropy loss function formula is written in equation 3.3 [64]. Where, N denotes number of samples, C is the number of classes, y_{ic} is binary indicator (returns 0 or 1) if class c is correct classification for sample i, p_{ic} is predicted probability whether sample i belongs to class c.

$$L = -\sum_{i=1}^{N}\sum_{c=1}^{C} y_{iC} \log{(p_{ic})} \tag{3.3}$$

Discussing the table 3.3, the layer column, which is followed by two columns, lists the kind of layers utilized in this model. On the other hand, the batch size, spatial dimensions, and number of filters or channels generated by the convolution are explained by the output shape column and its parameters. The model extracts feature maps by applying convolutional (Conv2D) layers to the input image and then processing the feature maps with pooling (MaxPooling2D) and fully connected (Dense) layers. First, the convolution layer (Conv2D) with kernel size of 3x3 with activation function ReLu is applied to input image. The output generated is 126x126 spatial dimensions of feature maps, 32 channels. The resulting feature maps are then down sampled by next max pooling (MaxPooling2D) layer with pool size of 2x2. Resulting output ("None", 63,63,32) will reduce the spatial dimensions in half. Along with that, the second convolutional layer uses 64 filters of size 3x3 for the refinement of feature extraction process as well. Thus, it produces the output shape of (None, 61, 61, 64). This layer is succeeded by the next max pooling layer, which effectively diminishes the spatial dimensions, thus resulting in an output shape of (None, 30, 30, 64). These layers jointly give rise to spatial refinement of the images by reducing the spatial feature size and increasing the depth of feature patches. When the convolutional and pooling layers are over, the network uses a flatten layer which is turned into a 1D vector shape of (None, 57600) in order to be an input to the fully connected layers. The last hidden layer, having the same number of neurons as the number of classes or similarity categories, finally returns the classification result in an output 12 neuron dense layer by using activation function SoftMax. SoftMax function is employed to produce probabilities over all the classes from the obtained logits. This change enables each output neuron to become the measure of the input instance belonging to a specific class, which makes the neural network ideal for use in multi-classification problems. Moreover, total parameters used in CNN model are 7393293 in which all the parameters are trainable and non-trainable parameters are zero.

Table 3.3: Layer detail of CNN model

Layer (type)	Output Shape	Parameters
Conv2D	(None,126,126,32)	320
MaxPooling2D	(None,63,63,32)	0
Conv2D	(None,61,61,64)	18496
MaxPooling2D	(None,30,30,64)	0
Flatten	(None,57600)	0
Dense	(None,128)	7372928

Dense	(None,12)	1548

3.5.4 Classifier

After the implementation of siamese network, next step is to classify the classes. K-nearest neighbor (KNN), support vector machine (SVM), random forest and decision tree have been used as the classifiers for our proposed model. The KNN algorithm is a basic machine learning method. The general concept is to find the distance between point A and every other point, eliminate the k points that are closest to the nearest point, and then count the k points that are part of the classification. Point A is included in the classification if it has the greatest proportion. We have used 5 neighbors in the algorithm. The majority voting concept is used by K-nearest neighbor classifiers to determine the class of a given data point. The classes of the five closest points are looked at if k is set to 5. Predictions are made based on the dominating class. In a similar vein, K-nearest neighbor regression uses the average of the five closest locations [55].

SVM is a supervised machine learning technique that is used for regression and classification problems. The classifier's main objective is to identify the hyperplane that divides the input data into the appropriate classes the best. The margin between the classes is maximized by selecting the hyperplane. What helps determine the position of the hyperplane are the data points that are closest to it, or support vectors. In order to define the margin and the decision border, these are essential. In high-dimensional spaces, SVM performs well. The kernel trick makes it versatile, enabling it to manage nonlinear interactions. Additionally, since it only uses a portion of the training set, it is memory-efficient.

An ensemble learning method called Random Forest is used to solve regression and classification issues. A machine learning technique called ensemble learning uses several models to solve a single issue in order to increase accuracy. Specifically, ensemble classification uses several classifiers to get findings that are more accurate than those of a single classifier. Put another way, combining several classifiers reduces variation and could lead to more dependable findings, particularly when dealing with unstable classifiers [61]. Each tree of decision tree in the forest partitions nodes with some randomly selected features after training on randomly selected samples from the training data. This is so because of their diversity and random nature; a collection of decision trees will produce predictions that are more dependable and accurate than single trees. Every tree in the forest votes for the class of its kind and the majority is selected as the final decision. The merging of the advantages of various models through an ensemble approach yields the ensemble's high accuracy, robustness against overfitting and the capacity to handle large datasets with increased dimensionality.

One of the widely used techniques, which can be found in many domains such as machine learning, image processing, data mining and pattern recognition is the Decision Tree. Every test in the DT includes a comparison of a numerical feature to a threshold value. The DT is a sequential model that effectively and cohesively combines these simple tests. In contrast to the quantitative weights in neural network, it is much easier to lay the conceptual principles. DT is used for grouping mostly. Each of the trees is consist of nodes and branches. Every subset will define a value that is acceptable by the node, and each node represents features in a category of data that needs to be labeled. Decision trees have been applied in many areas due to their simplicity of analysis and ability to produce accurate results for different data types [62].

3.6 Simulation Framework

Google colab has been used as network simulator to evaluate the performance of the proposed network. Backend is handled by tenserflow. Google offers a free cloud-based tool called google colab, sometimes known as Collaboratory, which enables users to write and run python code together in a Jupyter Notebook environment. Google collaboratory notebook is a virtual environment that gives users access to free GPU resources in Python. It is intended to make machine learning and data science activities easier. Users no longer need to set up and configure their own development environment because Colab runs in the cloud. It is therefore practical for rapid coding and teamwork. However, a number of well-known python libraries for data analysis, machine learning, and visualization including TensorFlow, PyTorch, Matplotlib, and others come pre-installed with colab. We uploaded our dataset on google drive and mount the drive to google colab to access data easily. Moreover, hyper parameters for Siamese network are set as; image dimension (128,128,1), batch size is 16, optimizer applied is adam, epoch is 300, learning rate is 0.0001 and loss function is triplet loss. Similarly, hyper parameters for CNN model are set as; batch size is 32, epoch is 10, validation-split is 0.1 and loss function applied is categorical cross-entropy.

This AI-based ship detection system can be deployed in various marine time environments, such as naval bases, commercial shipping lanes, and port authorities, to enhance security, monitor traffic, and assist in identifying various incoming vessels. It can also be used in marine conservation efforts to track and study the movements of different types of ships in ecologically sensitive areas, helping to prevent collisions with marine life or reduce underwater noise pollution. Additionally, it could be implemented in offshore oil and gas platforms to monitor nearby ship activity, ensuring operational safety and compliance with maritime regulations.

Depending on the deployment environment, i.e., whether the model will run on an edge device (e.g., a specialized microcontroller, Raspberry Pi, or NVIDIA Jetson) near the hydrophones or on a remote server. The choice depends on the computational power required and the latency constraints. real-time inference capabilities can be implemented so that the captured audio can be immediately processed by the proposed AI model. This might involve running a small software application on the edge device or server that continuously listens to the audio stream and feeds it to the model.

3.6.1 Performance Metrics

Underwater classification of ship engine is significant and becoming the hot research topic. As the technology is advancing, new techniques are also developing to make robust systems. Globally, underwater recognition systems have become a significant subset of general-purpose marine essential technologies, operating in numerous nations. Underwater target recognition has extensive application in several domains such as underwater defense and marine geological investigation. The major objectives of the research have been to improve the hydroacoustic signal processing and feature extraction techniques, optimize the sonar signal collection capability, and increase the accuracy of the system target recognition. A generation of hydroacoustic scientists has been plagued by low efficiency and accuracy due to the fact that the traditional underwater target recognition system only performs simple processing of the sonar collection signal. The sonar operator then judges the recognition of the characteristics of the hydroacoustic signal based on his personal experience [17]. The effect of accuracy will be seen by the proposed siamese network, which operates based on similarity. The sample to be compared will be further away from the sample of a different class.

Equation 3.4 [3] explains that the number of ship-engines correctly classified into the class to which they belong is represented by the true positives (TP), and the number of ship-engine correctly classified as not corresponding to a class is represented by the true negatives (TN). Furthermore, the number of ship-engine that are mistakenly identified as belonging to a class is represented by the false positives (FP), while the number of ship-engine that are improperly labelled as not belonging to a class is represented by the false negatives (FN).

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{3.4}$$

True-positives being the only correctly categorized windows of each class under investigation are normally used to calculate precision and recall for each class. The exactness of a prediction is a measure of how precisely it predicts for the particular category as mentioned in equation 3.5 [3].

$$\text{Precision} = \frac{TP}{TP+FP} \tag{3.5}$$

It determines the recall which is the percentage of true positives as mentioned in equation 3.6 [3].

$$\text{Recall} = \frac{TP}{TP+FN} \tag{3.6}$$

The harmonic mean is utilized in order to receive the F1-score for the precision and recall as mentioned in equation 3.7 [3].

$$\text{F1-score} = 2 \times \left(\frac{Precision \times Recall}{Precision + Recall}\right) \tag{3.7}$$

Moreover, plotting the true positive rate (TPR) vs the false positive rate (FPR) creates a ROC curve. The fraction of observations that were accurately predicted to be positive out of all positive observations is known as the true positive rate (TP/(TP + FN)). In a similar vein, the false positive rate (FP/(TN + FP)) represents the fraction of observations that are mistakenly anticipated to be positive out of all negative observations. For instance, in diagnostic testing, the true positive rate is the proportion of individuals who are accurately diagnosed as having the disease under investigation.

3.6.2 Underwater Methods for Classification

Underwater classification methods of deep learning are widely used. While acoustic signals are time-sequential signals, they typically provide additional information in their frequency domain. As a result, before feeding raw data into recognition models, it must be pre-processed and features extracted in order to minimize data dimensions and decrease noise [23]. Feature extraction is the cornerstone of underwater target recognition. Time-domain feature extraction, spectrum estimation methods, time-frequency analysis, and other approaches are the main feature extraction techniques used by researchers. As the submerged acoustic datasets grow, the first feature extraction techniques become less and less efficient.

Thus, developing novel techniques for underwater target recognition is crucial. Traditional underwater techniques focus on data augmentation [11] to extend the dataset. Moreover, developed integration of multimodal feature extraction techniques which leads to high complexity, cost and computational cost [7]. However, we proposed Siamese networks as feature extraction technique that are based on metric learning and are capable of comparing two instances in order to determine how similar they are. Siamese networks, in contrast to conventional machine learning techniques, are able to identify new targets from small amounts of data after being taught on an extensive number of previous samples [58].

3.7 Proposed Siamese Network

Proposed Siamese Network has one anchor or reference image that has been taken randomly from the dataset as an input. Second image is positive example of anchor image; positive image means image from same class as anchor image. Third image is negative example of anchor image; negative image means image from different class as anchor image. The aim of siamese network is to minimize the distance between anchor and positive image and maximize the distance between anchor and negative image as shown in figure 3.6.

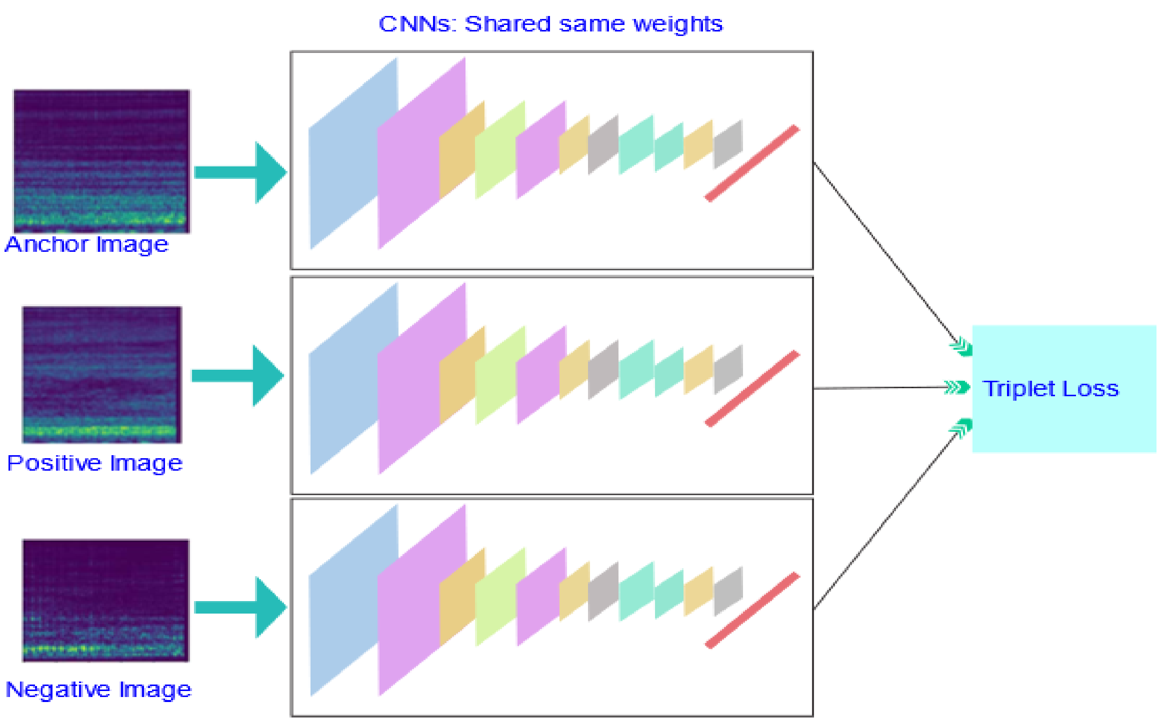

Figure 3.6: Siamese Tripplet Network

Applications in environmental monitoring, animal or object tracking [59], marine resource management, and maritime surveillance are all highly promising for our suggested methodology. Through precise categorization of ship-engine noises in submerged settings, our method can help with illicit activity detection, abnormal activity recognition, and maritime law enforcement. Siamese networks are also well-suited for deployment on fixed sensor arrays or autonomous underwater vehicles (AUVs) due to their scalability and flexibility, which allows for real-time monitoring and surveillance of maritime traffic. Moreover, precise temperature structure restoration and multi-view feature extraction are two more applications for the siamese network [60]. In summary, our study advances the current understanding of underwater acoustic signal processing and categorization methods, which has applications in improving maritime security, safeguarding marine environments, and encouraging sustainable marine development.

A Siamese network is designed to learn feature representations by comparing pairs of inputs, typically for tasks like similarity learning or matching. Two identical subnetworks with the same weights that analyse two input samples independently are trained as part of the Siamese network's feature selection process. Through the use of layers such as convolutional, pooling, and fully connected layers, which convert the inputs into feature vectors, the network retrieves feature from various sources. To find the degree of similarity between the inputs, these vectors are then compared, frequently using a distance metric such as Euclidean distance. High-dimensional representations of the key patterns and relationships found in the data are usually the features extracted by a Siamese network; these representations emphasise certain attributes such as shape, texture, or other pertinent properties that distinguish or correlate the input.

3.8 Summary

This chapter addresses the issues in existing underwater classification methods and provides the proposed solution. Dataset description and proposed model is also explained. Moreover, the operational framework provides the flow the of overall work with detailed steps. Furthermore, this section provides the working of Siamese network along with CNN model, classifiers and performance metrices.

CHAPTER NO 4

PERFORMANCE EVALUATION OF UNDERWATER SHIP-ENGINE CLASSIFICATION USING SIAMESE NETWORK (USCSN)

4.1 Overview

Chapter 4 presents a thorough analysis of the simulation findings and conclusions to underwater ship-engine classification using siamese network. The comparison of two classifiers has been discussed. This chapter discussed the results of using various performance measurements, including confusion matrix and accuracy. In addition, an explanation, several diagrams, and a comparison of the classifiers are given. SNE-plot that illustrates how accuracy affects the suggested model. The genuine positive rate above the thresholds is explained by the ROC curve testing, though.

4.2 Results and Analysis:

To evaluate the performance of proposed model, following matrices are used; accuracy, precision, recall, f1-score. Additionally, comparison of siamese network and CNN network have been reviewed through confusion matrix. Moreover, ROC testing, quantile plot has been plotted. ShipEars dataset discussed in pervious chapter is used to analyze results.

Moreover, the parameter setting of siamese network is shown in table 4.1 below in which the batch size, image dimension, learning rate are shown.

Table 4.1: Hyperparameter setting of Siamese Network

Hyperparameter	Value
Input shape	(128,128,1)
Batch size	16
Epoch	300
Learning rate	0.0001
Optimizer	Adam
Loss	Triplet loss

The parameter setting of CNN model without using siamese network is shown in table 4.2.

Table 4.2: Hyperparameter setting of CNN model

Hyperparameter	Value
Batch size	32
Epoch	10
Validation split	0.1
Optimizer	Adam
Loss	Categorical Cross entropy

4.3 Comparison of Classifiers:

In SVM, the linear Kernel works fine with the data that can be segmented into straight line or with linearly separable data itself. It is less power-thirsty and compact and also performs well on the datasets with big dimensions and sparsity, as in the case of text classification of high-dimensional feature vectors. The classification report of linear kernel is explained in table 4.3. While linear kernel management achieved a general accuracy of 89%, it showed good class separation performance. Most of the classes have high precision recall and also a good F1-scores, but there is a problem with classes 1 and 7 as they have lower performance, this means that these special categories are difficult to be different from others. This model works well for scenarios involving easier and linearly separable conditions but tends to face certain challenges which emanate from more intricate boundaries.

Table 4.3: Classification report of linear kernel SVM

Classes	Precision	Recall	F1-score
1	0.67	0.57	0.62
2	1.00	0.71	0.83
3	0.75	0.86	0.80
4	1.00	1.00	1.00
5	1.00	1.00	1.00
6	0.78	1.00	0.88
7	0.62	0.76	0.67
8	1.00	1.00	1.00
9	1.00	1.00	1.00
10	1.00	1.00	1.00
11	1.00	1.00	0.97
12	1.00	1.00	1.00

A polynomial function that fits the relationship between classes is a sufficient condition for applying the polynomial kernel to periodically create polynomial decision boundaries. It is the least versatile kernel among all kernels, but it has a higher level of flexibility than the linear kernel because of the degree parameter, which facilitates different forms of complexity. When the polynomial relationship is better at describing the nonlinear relationships between features in the data set, then the polynomial kernel is very efficient. The classification report of polynomial kernel is explained in table 4.4. With the correction 90 percent for overall polynomial kernel the improvement in performance of the classifier is fair, as compared to the linear kernel. Referring to almost every class, it is very good for remembering and details, providing advanced non-linear relationship recognition.

Table 4.4: Classification report of polynomial kernel SVM

Classes	Precision	Recall	F1-score
1	0.62	0.71	0.67
2	1.00	0.86	0.92
3	0.88	1.00	0.96
4	1.00	1.00	1.00
5	1.00	1.00	1.00
6	0.86	0.86	0.86
7	0.67	0.57	0.62
8	1.00	1.00	1.00
9	1.00	1.00	1.00
10	1.00	1.00	1.00
11	0.94	1.00	1.00
12	0.86	0.86	0.86

On the contrary, the radial basis function (RBF) kernel is best in this instance when the decision boundary is non-linear. Through the mapping of the input data into a higher-dimensional space, the RBF kernel in an effective manner is capable of detecting the nonparametric and complicated patterns. It is the best fit for the data sets with intricate interdependencies that cannot be reflected by the limits of a straight-line boundary due to its flexibility. The technology is comprehensively used in the scenarios where the data depicts complex patterns like image classification and bioinformatics. The classification report of radial basis kernel is explained in table 4.5. RBF also reaches 90 % of accuracy, revealing that non-linear decision boundaries are effectively dealt with by this kernel. As far as most classes are concerned this model performs quite reliably; however, in the case of classes 6 and 7 it displays poor precision and recall. The RBF kernel, that possesses the feature of accurately capturing the intricate patterns, provides with consistent performance across the dataset as a whole.

Table 4.5: Classification report of radial basis kernel SVM

Classes	Precision	Recall	F1-score
1	0.88	1.00	0.93
2	1.00	0.86	0.92
3	0.86	0.86	0.86
4	1.00	0.86	0.92
5	1.00	1.00	1.00
6	0.58	1.00	0.74
7	0.90	0.97	0.93
8	1.00	0.57	0.73
9	1.00	1.00	1.00
10	1.00	1.00	1.00
11	0.88	1.00	0.93
12	1.00	0.71	0.83

An overall accuracy of 89% in figure 4.4 and management of most classes in the K-NN classifier played well in this context. Although the concept of a recall was assumed by people, its implication is not desired. The model is off target with accuracy of 57 for class 7, foregrounding the prospect that it is extremely difficult to determine all instances originating from a given class. The classification report of KNN is explained in table 4.6.

Table 4.6: Classification report of KNN

Classes	Precision	Recall	F1-score
1	0.86	0.86	0.86
2	0.88	1.00	0.93
3	0.86	0.86	0.86
4	0.71	0.71	0.71
5	1.00	1.00	1.00
6	1.00	1.00	1.00
7	0.80	0.57	0.67
8	1.00	1.00	1.00

9	0.70	0.86	0.80
10	1.00	0.80	0.93
11	0.94	0.80	0.93
12	1.00	1.00	1.00

The classification report of random forest is explained in table 4.7. The report on random forest classifier shows good performance, the overall accuracy being 86 percent and excellent precision and recall for many others classes. It also should be noted that it's weak, class 12, and a poor F1-score of 0. 55 which suggests low performance, for example: not all instances of this class may be labeled properly.

Table 4.7: Classification report of Random Forest

Classes	Precision	Recall	F1-score
1	1.00	0.71	0.83
2	0.78	1.00	0.88
3	1.00	0.86	0.92
4	0.83	0.71	0.77
5	1.00	1.00	1.00
6	1.00	0.71	0.71
7	0.71	1.00	0.82
8	1.00	1.00	1.00
9	1.00	0.86	1.00
10	0.88	1.00	0.93
11	0.78	1.00	0.88
12	0.75	0.43	0.55

The classification report of decision tree is explained in table 4.8. However, the decision tree classifier, basically has lower accuracies, recall, and precision for most classes apart from overall accuracy of 52%. This illustrates the power of the grouping method, for example, random forest but not the decision tree for more accurate classification on its own. This is also a symptom of poor complexity performance and of increased uncertainty or randomness of predictions.

Table 4.8: Classification report of Decision Tree

Classes	Precision	Recall	F1-score
1	0.83	0.71	0.77
2	0.57	0.57	0.57
3	0.33	0.14	0.20
4	0.25	0.14	0.18
5	0.99	0.71	0.83
6	0.38	0.86	0.40
7	0.50	0.29	0.59
8	0.55	0.86	0.67
9	0.40	0.71	0.33
10	0.50	0.14	0.63
11	0.45	0.71	0.56
12	0.50	0.14	0.22

Moreover the, training and testing results of all classifiers are shown in table 4.9. As our dataset is non-linear so linear kernel couldn't accurately classify the classes and might confuse with other classes. Radial basis kernel and polynomial kernel on the other hand reached expected accuracy. The results shows that KNN classifier performs well with 95.857%.

Table 4.9: Classifiers accuracy in training and testing

Classifier	Kernel	Accuracy Training	Accuracy Testing
SVM	Linear	98.276%	89.286%
SVM	Polynomial	94.540%	90.476%
SVM	Radial basis	99.425%	90.476%
KNN	**-**	**96.839%**	**95.857 %**

RF	-	100%	86.00%
DT	-	100%	52.00%

4.4 Comparison of Siamese Network with CNN model:

The accuracy recorded of the proposed siamese network by using the confusion matrix is 96.4167% as shown in figure 4.1. True labels are shown in the diagonal by dark blue color. Light blue color or white color shows false label. “1” written in dark blue color represents that the predicted label shows 100% accuracy with true label. For instance, dredger in predicted and true tables marked as “1”, which shows that our model perfectly classifies the true class of dredger. However, tugboat marked as “0.71” in diagonal means that our model correctly classifies 71% for tugboat and confuses 29% tugboat with pilotboat. Similarly, our model correctly classifies 86% for ambient noise and confuses 14% ambient noise with yacht.

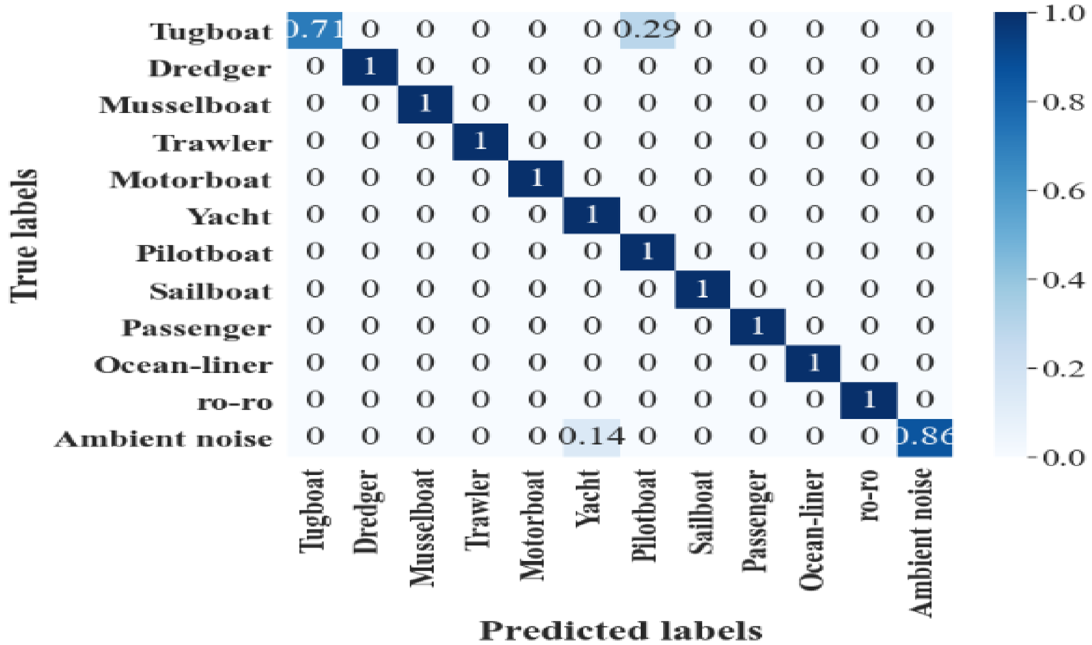

Figure 4.1: Confusion Matrix for Siamese Network

The accuracy recorded of the CNN network by using the confusion matrix is 94.0476% as shown in figure 4.2. Labels marked close to 1 are close to accuracy. "1" written in diagonal in dark blue color represents that the predicted label shows 100% accuracy with true label. For instance, mussel boat in predicted and true tables marked as "1", which shows that our model perfectly classifies the true class of mussel boat. However, CNN model correctly classifies 57% for ambient noise and confuses 43% ambient noise with yacht. Our model out performs the CNN network by 2.3691%.

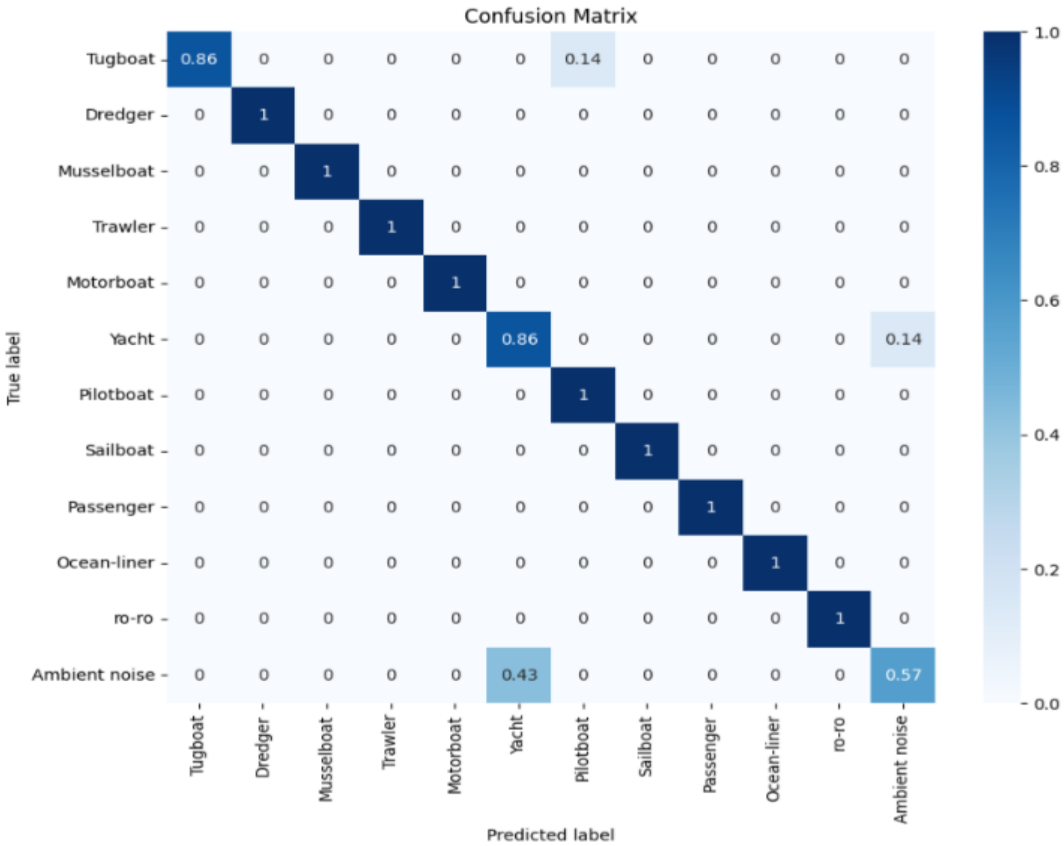

Figure 4.2: Confusion Matrix for CNN

4.5 Receiver Operator Characteristic (ROC) Curve Testing

The ROC curve is a graphical plot that illustrates how well the classifiers can diagnose problems. False positives, or misclassified samples, divided by total negatives, or false positives plus true negatives, is known as the false positive rate. Put simply, the number of times the response is recorded as yes when it is actually no. The ratio of accurately identified samples, or true positives, to total positive samples, or true positives plus false negatives, is known as the true positive rate. Put another way, the number of times the response is reported as yes when it is a yes.

The figure 4.3 illustrates the true positive rate close to 1.0 and false positive rate below the threshold 0.5 against threshold. We might repeatedly assess a logistic regression model with various classification criteria in order to calculate the points in a ROC curve. ROC curve of class 1 to 12, true positive rate is high and false positive rate is low for our proposed model.

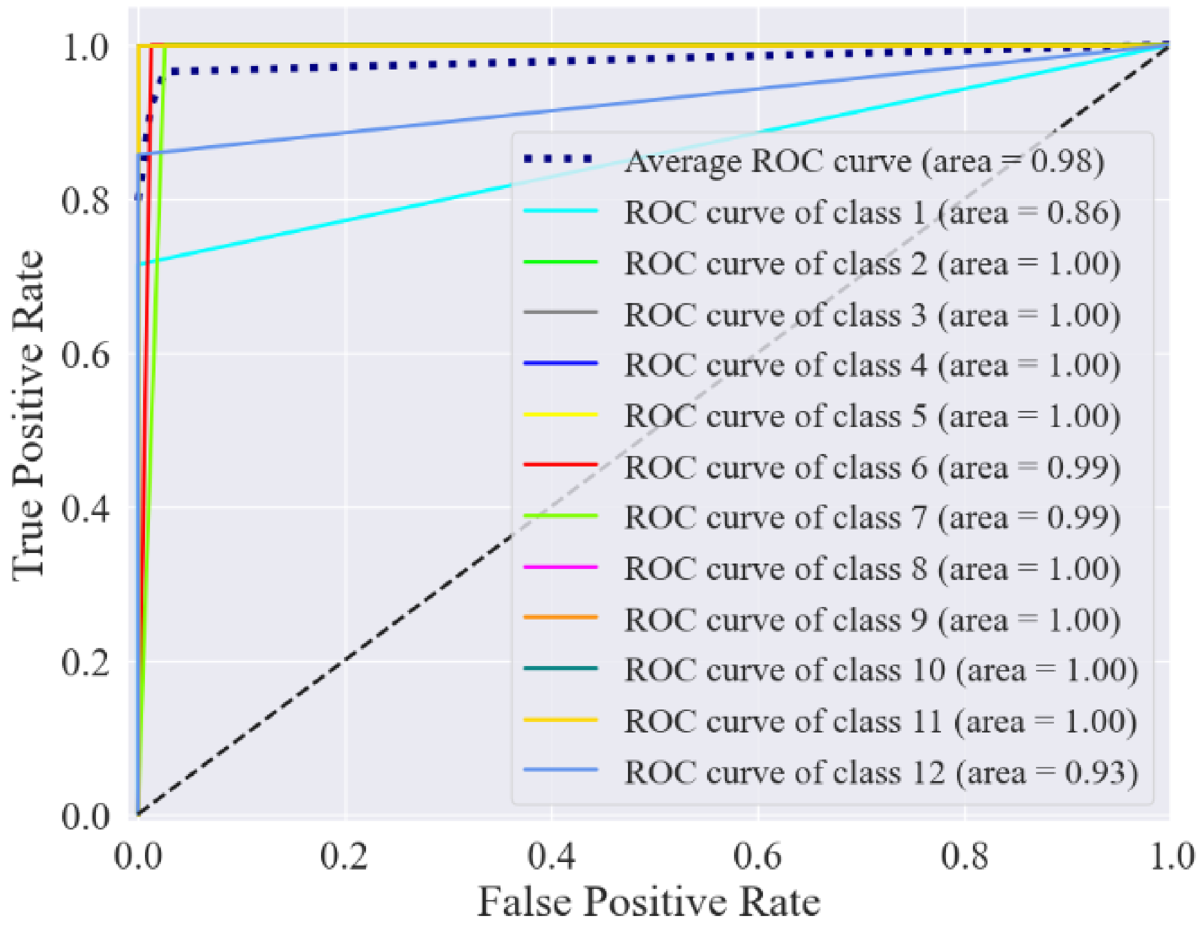

Figure 4.3: ROC curve testing of the proposed model

Visually comparing the quantiles of an observed dataset to those of a theoretical distribution is necessary to interpret a Q-Q display. It can evaluate how well the observed dataset fits the expected distribution and spot deviations from normalcy or other assumptions by analyzing the linearity of the points and deviations from the diagonal line. Figure 4.4 shows the examines the arrangement of the points on the Q-Q plot. A straight line connecting the points indicates that the observed dataset closely resembles the theoretical distribution. The quantiles of the observed dataset are represented on the y-axis of the Q-Q plot, whilst the quantiles of the theoretical distribution, such as the standard normal distribution, are represented on the x-axis. Although the ends of the Q-Q plot frequently begin to diverge from the straight line, typically distributed data appears as roughly a straight line on a Q-Q plot. The points on the Q-Q plot will sit along the diagonal line (y = x) if the observed dataset exactly matches the theoretical distribution. This would suggest a tight match between the two distributions.

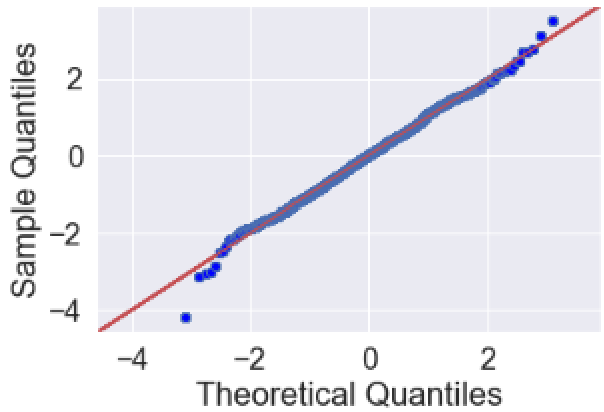

Figure 4.4: Q-Q plot

While both Convolutional Neural Networks (CNNs) and Support Vector Machines (SVMs) are effective machine learning models, their architectures and uses are very different. CNNs are deep learning models created especially to handle data that resembles a grid, like pictures. Their effectiveness for tasks like image and audio identification stems from their ability to automatically learn spatial hierarchies of features through layers of convolutional filters. On the other hand, SVMs are conventional machine learning algorithms that operate by determining the best hyperplane in a high-dimensional space to divide data points belonging to various classes. SVMs are good at binary classification and perform well on small to medium-sized datasets, but CNNs perform better on large-scale jobs where feature extraction is important since they are better at handling complicated, high-dimensional data [65].

4.6 Summary

This chapter discusses the simulation results and proposed technique's performance. Modified shipears dataset is used for evaluation. A comparative analysis of the suggested strategy with pertinent schemes has also been done, taking performance metrics into account. The results demonstrate that the suggested model outperformed the terms for every performance indicator.

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 Overview

This chapter discusses the conclusion and future work of the research work. The main aim of this research was to introduce the new technique for classification of ship-engine. The proposed siamese network acts as the feature extraction technique. It performs well for small dataset by creating triplets from each class. On the similarity basis, siamese network uses triplet loss function to create more distance between samples of different class and reduce distance between sample of same class. Additionally, four classifiers, SVM, k-NN, RF and DT were used to classify the classes based on results generated by siamese network. Evaluation matrices and plots demonstrate how well the suggested network operates.

5.2 Summary and Contribution

Our research's primary contribution is the suggestion to use Siamese networks for the underwater classification of audio signals from ship engines. Siamese networks present a potentially effective way to tackle the problems related to ship-engine sound classification in underwater settings. Siamese networks are naturally suited for jobs with few labelled samples, unlike traditional classification techniques that require labelled training data for every type of ship. They can also efficiently learn from tiny datasets using one-shot or few-shot learning paradigms. The twin network design of Siamese networks is typified by the simultaneous training of two identical subnetworks on pairs of input samples. Each input sample is mapped by the Siamese network during training into a high-dimensional embedding space, where samples belonging to the same class are densely packed together and samples belonging to different classes are separated by a greater margin. Because of this, even with a small amount of labelled data, the Siamese network can discriminatively classify ship-engine audio signals by successfully capturing their inherent similarities and variances. Siamese networks can also be trained with contrastive or triplet loss functions to maximize the embedding space and push different samples farther apart and comparable samples closer together.

The proposed siamese network outperforms the other methods by 96.4167% accuracy. Siamese networks perform greatly better in view of learning correlations or similarities between multiple classes of data. Whereas traditional CNNs are mostly used for detection as well as classification of objects, Siamese networks are especially suited for matching and comparison of object pairs on the basis of their similarity. One of the main advantages of Siamese networks is the ability to learn from few labelled datasets and the less available of pairings than typical CNNs which requires massive labelled datasets. Some of them are particularly helpful in the circumstances that labelled data is pricey or in shortage. Siamese networks apply well in face recognition systems as well as signature verification, and image comparison tasks because of their capability to adapt to changing input data and increase their accuracy during generalization without any modifications.

Not only do they have the possibility of using various tasks such as contrastive or triplet loss for loss functions, they also maximize the similarity between similar pairings, minimizing the similarity between different pairs, resulting in discrete embeddings. To conclude, Siamese networks are of great importance for applications where important task is learning association or pair between data items.

However, in order to thoroughly assess the effectiveness of the suggested Siamese network model for underwater ship-engine sound classification, we want to compare various classification techniques, including Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), Random Forest and Decision Tree. We may evaluate the Siamese network's efficacy in managing the inherent difficulties of underwater acoustic signal classification, such as a lack of labelled data, noise, and environmental variability, by comparing its performance through different classifiers.

We can understand the relative performance in terms of classification accuracy, robustness to noise, scalability, and generalization ability by comparing these classifiers' performance with that of the Siamese network. By choosing SVMs (Support Vector Machines) as the classification algorithm for underwater ship-engine identification, robust decision boundaries are created in a high-dimensional space making it possible to differentiate between the patterns of different types of ship engines, even in the context of high background noise and acoustic variability. KNN (K-Nearest Neighbors) has the advantages of simplicity and high performance when dealing with data that is well-clustered like this, which make it a good choice for ship-engine sound classification that is based on the proximity to the known examples. Decision Trees are a valuable tool because they can be easily interpreted and analyzed, thus the analyst is able to see the decision-making process as well as to identify the specific acoustic features that distinguish one engine type from another. The Random Forest approach provides class enhancement which results from ensemble learning being a combination of many decision trees to improve accuracy, reduce overfitting and therefore facilitate in the proper management of complex and noisy underwater acoustic data sets. For underwater ship-engine classification jobs, this comparison analysis will offer useful information for choosing the best classification method based on particular needs and limitations. The results shows that k-NN classifier with 95.857% performs well than other classifiers.

5.3 Future Work

One of the drawbacks of triplet networks in the categorization of the underwater ship engine is their dependence on pre-selected triplets that are used for training but correctly curated samples in noisy environments are quite hard to find. Besides, a small size of dataset gives rise to another problem namely, how many triplet pairs will be necessary to train suitably embedding, and in the field of underwater acoustics, it is often difficult to collect sufficient labeled data. Not only that, issues like the network being vulnerable to adversarial attacks are also raised, which means that small changes in the input data could lead to classification errors lowering the reliability of the system in safety critical scenarios. Such limitations emphasize the necessity of the careful preparations of the dataset, careful selection of the triplet vector, and the efficient use of the security measures to make this domain approach closer to becoming reality.

Future work in the domain of underwater ship-engine classification by Siamese networks may concentrate on many aspects to propel the frontiers and tackle current obstacles. Here are a few possible avenues for further investigation:

- **Data Enrichment and Generated Synthetic Data:** Examine methods for creating artificial ship-engine sound to enhance current datasets. This can enhance the resilience of Siamese network models to changes in ship types, engine settings, and environmental factors, as well as help address problems associated with sparsely labelled data.
- **Integration of Sensors with Multi-Modal Fusion:** Examine how to improve ship-engine classification performance by integrating several sensing modalities, such as optical, audio, and environmental sensor data. Create fusion strategies to increase classification accuracy and reliability by utilizing complimentary data from various sensor types.
- **Both Domain Adaptation and Transfer Learning:** To generalize Siamese network models across various underwater settings, ship types, and acoustic sensor configurations, investigate transfer learning and domain adaption strategies. Formulate plans for knowledge transfer from labelled source domains with limited labelled data to unlabeled target domains.
- **Deployment in Real-Time and Autonomous Function:** Create and put into service real-time ship-engine classification systems that may be deployed autonomously on underwater sensor platforms, like stationary sensor arrays or autonomous underwater vehicles (AUVs). Provide hardware architectures and effective algorithms that are suited for embedded, low-power computing in submerged situations.
- **Finding anomalies and diagnosing faults:** Expand the capability of ship-engine classification frameworks to encompass anomaly detection and failure diagnosis. Provide algorithms that, when compared to typical auditory patterns, can identify anomalous engine behavior, mechanical issues, or environmental disruptions.
- **Deep Learning Frameworks and Model Enhancement:** Examine cutting-edge regularization strategies, optimization tactics, and deep learning architectures specifically designed for underwater ship-engine classification applications. Examine methods for hardware acceleration, quantization, and model compression to facilitate effective deployment on underwater platforms with limited resources.

- **Adversarial Security and Robustness:** Examine how resilient Siamese network models are to security risks and hostile attacks in submerged environments. Provide defenses against hostile perturbations to lessen their effects on classification performance and protect the integrity and dependability of ship-engine classification systems.
- **Explainability and Interpretability:** Improve the ship-engine classification models' interpretability and explainability to shed light on the features that discriminate and the decision-making process. To increase trust and comprehension of the categorization, look into methods for displaying and analyzing the learnt representations of Siamese network models.
- **Validation studies and field trials:** To assess how well Siamese network-based ship-engine classification systems function in actual underwater conditions, conduct field tests and validation investigations. Assess the practical usability and efficacy of the suggested solutions in cooperation with government agencies, research institutes, and stakeholders in the maritime industry.

In order to provide dependable and effective methods for identifying and monitoring ships in the underwater environment, future research in the field of underwater classification of ship-engine using Siamese networks can advance marine science, underwater technology, maritime operations, and maritime security by addressing these research directions. Moreover, different datasets can be used in future for better comparison.

REFERENCES

[1] Jiang, Zhe, Chen Zhao, and Haiyan Wang. "Classification of Underwater Target Based on S-ResNet and Modified DCGAN Models." *Sensors* 22, no. 6 (2022): 2293.

[2] Han, Xing Cheng, Chenxi Ren, Liming Wang, and Yunjiao Bai. 2022. "Underwater acoustic target recognition method based on a joint neural network." *Plus One* 15.

[3] Vaz, Guilherme, Alexandre Correia, Miguel Vicente, Joao Sousa, Erica Cruz, and Benedicte Dommergues. "Marine Acoustic Signature Recognition Using Convolutional Neural Networks." *Available at SSRN 4119910*.

[4] Hong, Feng, Chengwei Liu, Lijuan Guo, Feng Chen, and Haihong Feng. "Underwater acoustic target recognition with a residual network and the optimized feature extraction method." Applied Sciences 11, no. 4 (2021): 1442.

[5] Santos-Domínguez, David, Soledad Torres-Guijarro, Antonio Cardenal-López, and Antonio Pena-Gimenez. "ShipsEar: An underwater vessel noise database." *Applied Acoustics* 113 (2016): 64-69.

[6] Hu, Gang, Kejun Wang, Yuan Peng, Mengran Qiu, Jianfei Shi, and Liangliang Liu. "Deep learning methods for underwater target feature extraction and recognition." *Computational intelligence and neuroscience* 2018 (2018).

[7] Luo, Xinwei, Yulin Feng, and Minghong Zhang. "An underwater acoustic target recognition method based on combined feature with automatic coding and reconstruction." *IEEE Access* 9 (2021): 63841-63854.

[8] Doan, Van-Sang, Thien Huynh-The, and Dong-Seong Kim. "Underwater acoustic target classification based on dense convolutional neural network." IEEE Geoscience and Remote Sensing Letters (2020).

[9] Xiao, Xu, Wenbo Wang, Qunyan Ren, Peter Gerstoft, and Li Ma. "Underwater acoustic target recognition using attention-based deep neural network." JASA Express Letters 1, no. 10 (2021): 106001.

[10] Yang, Honghui, Kaifeng Zheng, and Junhao Li. "Open set recognition of underwater acoustic targets based on GRU-CAE collaborative deep learning network." Applied Acoustics 193 (2022): 108774.

[11] Luo, Xinwei, Minghong Zhang, Ting Liu, Ming Huang, and Xiaogang Xu. 2021. "An Underwater Acoustic Target Recognition Method Based on Spectrograms with Different Resolutions." *Journal of Marine and Science and Engineering* 14.

[12] Ayvaz, Uğur, Hüseyin Gürüler, Faheem Khan, Naveed Ahmed, and Taegkeun Whangbo. "Automatic speaker recognition using mel-frequency cepstral coefficients through machine learning." (2022).

[13] Ma, Yanxin, Mengqi Liu, Yi Zhang, Bingbing Zhang, Ke Xu, Bo Zou, and Zhijian Huang. "Imbalanced underwater acoustic target recognition with trigonometric loss and attention mechanism convolutional network." Remote Sensing 14, no. 16 (2022): 4103.

[14] Honghui, Yang, Li Junhao, and Sheng Meiping. "Underwater acoustic target multi-attribute correlation perception method based on deep learning." Applied Acoustics 190 (2022): 108644.

[15] Zhu, Chenyang, Tomaso Gaggero, Nicholas C. Makris, and Purnima Ratilal. "Underwater Sound Characteristics of a Ship with Controllable Pitch Propeller." Journal of Marine Science and Engineering 10, no. 3 (2022): 328.

[16] PARLAK, Cevahir, and Yusuf ALTUN. "A novel filter bank design for speech emotion recognition." (2022).

[17] Nanni, Loris, Giovanni Minchio, Sheryl Brahnam, Gianluca Maguolo, and Alessandra Lumini. "Experiments of image classification using dissimilarity spaces built with siamese networks." Sensors 21, no. 5 (2021): 1573.

[18] El Boghdady, Nawal, Florian Langner, Etienne Gaudrain, Deniz Baskent, and Waldo Nogueira. "Effect of spectral contrast enhancement on speech-on-speech intelligibility and voice cue sensitivity in cochlear implant users." Ear and hearing 42, no. 2 (2021): 271-289.

[19] Parikh, Dishant, and Saurabh Sachdev. "Improving the efficiency of spectral features extraction by structuring the audio files." In 2020 IEEE-HYDCON, pp. 1-5. IEEE, 2020.

[20] Humphrey, Eric J., Taemin Cho, and Juan P. Bello. "Learning a robust tonnetz-space transform for automatic chord recognition." In 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 453-456. IEEE, 2012.

[21] Manocha, Pranay, Rohan Badlani, Anurag Kumar, Ankit Shah, Benjamin Elizalde, and Bhiksha Raj. "Content-based representations of audio using siamese

neural networks." In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3136-3140. IEEE, 2018.

[22] Hu, G., Wang, K., Peng, Y., Qiu, M., Shi, J., & Liu, L. (2018). Deep learning methods for underwater target feature extraction and recognition. *Computational intelligence and neuroscience*, *2018*.

[23] Luo, X., Chen, L., Zhou, H., & Cao, H. (2023). A Survey of Underwater Acoustic Target Recognition Methods Based on Machine Learning. *Journal of Marine Science and Engineering*, *11*(2), 384.

[24] Ashraf, H., Shah, B., Soomro, A. M., Halim, Z., & Shah, S. K. (2022). Ambient-noise free generation of clean underwater ship engine audios from hydrophones using generative adversarial networks. *Computers and Electrical Engineering*, *100*, 107970.

[25] Dong, X., & Shen, J. (2018). Triplet loss in siamese network for object tracking. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 459-474)

[26] Bhatt, Bhrugu, et al. "A Comparative Study of Transfer Learning Networks and Siamese Networks for Acute Lymphoblastic Leukemia (ALL) Diagnosis." *Journal of Student Research* 11.4 (2022).

[27] Solomon, E., Woubie, A., & Cios, K. J. (2022). UFace: An Unsupervised Deep Learning Face Verification System. *Electronics*, *11*(23), 3909.

[28] Li, J., Zhu, S., Gao, Y., Zhang, G., & Xu, Y. (2022). Change detection for high-resolution remote sensing images based on a multi-scale attention siamese network. *Remote Sensing*, *14*(14), 3464.

[29] Song, C., & Ji, S. (2022). Face Recognition Method Based on Siamese Networks Under Non-Restricted Conditions. *IEEE Access*, *10*, 40432-40444.

[30] Zhao, B., Sui, H., & Liu, J. (2023). Siam-DWENet: Flood inundation detection for SAR imagery using a cross-task transfer siamese network. *International Journal of Applied Earth Observation and Geoinformation*, *116*, 103132.

[31] Nanni, L., Minchio, G., Brahnam, S., Maguolo, G., & Lumini, A. (2021). Experiments of image classification using dissimilarity spaces built with siamese networks. *Sensors*, *21*(5), 1573.

[32] Agrawal, R., & Dixon, S. (2021, January). Learning frame similarity using siamese networks for audio-to-score alignment. In *2020 28th European Signal Processing Conference (EUSIPCO)* (pp. 141-145). IEEE.

[33] Abbas, A., & Shoaib, M. (2022). Kinship identification using age transformation and Siamese network. *PeerJ Computer Science*, *8*, e987.

[34] Nanni, L., Rigo, A., Lumini, A., & Brahnam, S. (2020). Spectrogram classification using dissimilarity space. *Applied Sciences*, *10*(12), 4176.

[35] Tian, S., Bai, D., Zhou, J., Fu, Y., & Chen, D. (2023). Few-shot learning for joint model in underwater acoustic target recognition. *Scientific Reports*, *13*(1), 17502.

[36] Cheng, X., & Zhang, H. (2021). Underwater target signal classification using the hybrid routing neural network. *Sensors*, *21*(23), 7799.

[37] Honka, T. (2019). One-shot Learning with Siamese Networks for Environmental Audio.

[38] Burguera, A. (2021). Lightweight underwater visual loop detection and classification using a Siamese convolutional neural network. *IFAC-PapersOnLine*, *54*(16), 410-415.

[39] Lee, S., & Kim, E. (2018). Multiple object tracking via feature pyramid siamese networks. *IEEE access*, *7*, 8181-8194.

[40] Zhang, Y., Lin, Q., Tang, H., & Li, Y. (2023). Research on Siamese Object Tracking Algorithm Based on Knowledge Distillation in Complex Marine Environment. *IEEE Access*.

[41] Dawkins, M., Prior, J., Lewis, B., Faillettaz, R., Banez, T., Salvi, M., ... & Hoogs, A. (2024). FishTrack23: An Ensemble Underwater Dataset for Multi-Object Tracking. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 7167-7176).

[42] Wu, X., Han, X., Zhang, Z., Wu, H., Yang, X., & Huang, H. (2023). A Hybrid Excitation Model Based Lightweight Siamese Network for Underwater Vehicle Object Tracking Missions. *Journal of Marine Science and Engineering*, *11*(6), 1127.

[43] Yin, W., Ye, Z., Peng, Y., & Liu, W. (2023, May). A review of visible single target tracking based on Siamese networks. In *2023 4th International Conference on Electronic Communication and Artificial Intelligence (ICECAI)* (pp. 282-289). IEEE.

[44] Lee, M. F. R., & Chen, Y. C. (2023). Artificial Intelligence Based Object Detection and Tracking for a Small Underwater Robot. *Processes*, *11*(2), 312.

[45] Zhang, J., Huang, H., Jin, X., Kuang, L. D., & Zhang, J. (2023). Siamese visual tracking based on criss-cross attention and improved head network. *Multimedia Tools and Applications*, 1-27.

[46] Chen, S., Chen, S., Shan, W., & Xu, Y. (2023, September). A self-supervised underwater image denoising method based on pseudo-siamese neural network. In *2023*

5th International Conference on Robotics and Computer Vision (ICRCV) (pp. 130-134). IEEE.

[47] Bertinetto, L., Valmadre, J., Henriques, J. F., Vedaldi, A., & Torr, P. H. (2016). Fully-convolutional siamese networks for object tracking. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II 14* (pp. 850-865). Springer International Publishing.

[48] He, A., Luo, C., Tian, X., & Zeng, W. (2018). A twofold siamese network for real-time object tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4834-4843).

[49] Guo, Q., Feng, W., Zhou, C., Huang, R., Wan, L., & Wang, S. (2017). Learning dynamic siamese network for visual object tracking. In *Proceedings of the IEEE international conference on computer vision* (pp. 1763-1771).

[50] Zhu, Z., Wang, Q., Li, B., Wu, W., Yan, J., & Hu, W. (2018). Distractor-aware siameses networks for visual object tracking. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 101-117).

[51] Shen, J., Tang, X., Dong, X., & Shao, L. (2019). Visual object tracking by hierarchical attention siamese network. *IEEE transactions on cybernetics*, *50*(7), 3068-3080.

[52] Zhang, H., Xing, W., Yang, Y., Li, Y., & Yuan, D. (2023). SiamST: Siamese network with spatio-temporal awareness for object tracking. *Information Sciences*, *634*, 122-139.

[53] Abbas, A. N., & Moser, D. (2021). Siamese Network Training using Artificial Triplets by Sampling and Image Transformation. *arXiv preprint arXiv:2106.07015*.

[54] Shi, D., Orouskhani, M., & Orouskhani, Y. (2021). A conditional Triplet loss for few-shot learning and its application to image co-segmentation. *Neural Networks*, *137*, 54-62.

[55] Lucchetti, G. P. (2020). Aplicação de machine learning para predição de manutenção de máquinas em empresa de transporte e logística de combustíveis.

[56] Dakhil, R. A., & Khayeat, A. R. H. (2022). Review On Deep Learning Technique For Underwater Object Detection. arXiv preprint arXiv:2209.10151.

[57] AlZoman, R. M., & Alenazi, M. J. (2021). A comparative study of traffic classification techniques for smart city networks. Sensors, 21(14), 4677.

[58] Cui, X., He, Z., Xue, Y., Tang, K., Zhu, P., & Han, J. (2024). Cross-Domain Contrastive Learning-Based Few-Shot Underwater Acoustic Target Recognition. Journal of Marine Science and Engineering, 12(2), 264.

[59] Mei, Y., Yan, N., Qin, H., Yang, T., & Chen, Y. (2024). SiamFCA: A new fish single object tracking method based on siamese network with coordinate attention in aquaculture. Computers and Electronics in Agriculture, 216, 108542.

[60] Zhang, S., Yang, Y., Xie, K., Gao, J., Zhang, Z., Niu, Q., ... & Jia, S. (2024). Spatial-Temporal Siamese Convolutional Neural Network for Subsurface Temperature Reconstruction. IEEE Transactions on Geoscience and Remote Sensing.

[61] Sheykhmousa, Mohammadreza, et al. "Support vector machine versus random forest for remote sensing image classification: A meta-analysis and systematic review." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 13 (2020): 6308-6325.

[62] Charbuty, B., & Abdulazeez, A. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. Journal of Applied Science and Technology Trends, 2(01), 20 - 28.

[63] Abbas, A. N., & Moser, D. (2021). Siamese Network Training using Artificial Triplets by Sampling and Image Transformation. arXiv preprint arXiv:2106.07015.

[64] Glorot, X., Bordes, A., & Bengio, Y. (2011). Deep Sparse Rectifier Neural Networks. Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS), 15, 315-323.

[65] Jędrzejczyk, A., Firek, K., & Rusek, J. (2022). Convolutional neural network and support vector machine for prediction of damage intensity to multi-storey prefabricated RC buildings. Energies, 15(13), 4736.