COMPARISON OF GOOGLE EARTH ENGINE-BASED MACHINE LEARNING CLASSIFIERS FOR MAPPING AQUACULTURE PONDS IN SUNGAI UDANG, PENANG

ARVINTH A/L RAJANDRAN

UNIVERSITI SAINS MALAYSIA

2023

COMPARISON OF GOOGLE EARTH ENGINE-BASED MACHINE LEARNING CLASSIFIERS FOR MAPPING AQUACULTURE PONDS IN SUNGAI UDANG, PENANG

by

ARVINTH A/L RAJANDRAN

Thesis submitted in fulfilment of the requirements
for the degree of
Master of Arts

February 2023

ACKNOWLEDGEMENT

I would like to take this opportunity to express my sincere gratitude to my supervisor, Associate Professor GS. Dr. Tan Mou Leong for his support, guidance and encouragement. With his prompt and useful advice, he had helped me to reach this finishing point. Then, I would like to thank my co-supervisor, Dato' Professor Dr. Narimah Samat for her assistance and guidance. They have helped tremendously especially in improving my work.

It is also my privilege to thank my colleague and friends, Zibeon bin Luhaim, Tew Yi Lin and Zeng Ju. They had given me their full support throughout this research and also a lifetime of unforgettable memories of their kindness, support and erudition. They kept showering me with motivation all the way to the end with their experiences and useful knowledge.

The research for this thesis was financially funded and supported by the Ministry of Higher Education of Malaysia under the Long-Term Research Grant Scheme (LRGS/1/2018/USM/01/1/5) (203/PHUMANITI/67215003). Finally, thanks to all parties and Universiti Sains Malaysia for their facilities and funding for this research.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

API	Application Programming Interface
CART	Classification and Regression Tree
FAO	Food and Agriculture Organization
GEE	Google Earth Engine
GDP	Gross Domestic Product
ha	Hectare
IDE	Integrated Development Environment
kg	Kilogram
LULC	Land Use Land Cover
MLC	Maximum Likelihood Classification
MNDWI	Modified Normalized Difference Water Index
MODIS	Moderate Resolution Imaging Spectroradiometer
NDBI	Normalized Difference Built-up Index
NDVI	normalized difference vegetation index
NGO	Non-governmental Organization
OBIA	Object-based image analysis
NIR	Near-Infrared
OA	Overall Accuracy
PA	Producer's Accuracy
RF	Random Forest
RGB	Red Green Blue
RS	Remote Sensing
SAR	Synthetic-aperture radar
SVM	Support Vector Machine

UA	User’s Accuracy

LIST OF APPENDICES

PERBANDINGAN PENGKELAS PEMBELAJARAN MESIN BERASASKAN *GOOGLE EARTH ENGINE* UNTUK PEMETAAN KOLAM AKUAKULTUR DI SUNGAI UDANG, PULAU PINANG

ABSTRAK

Penderiaan jauh merupakan salah satu cara yang paling sesuai untuk memantau dan mengurus aktiviti akuakultur. Pemetaan kolam akuakultur boleh dijalankan dengan lebih mudah dan cepat kerana kemajuan pesat dalam pengkomputeran awan, i.e., *Google Earth Engine* (GEE). Beberapa pengelas pembelajaran mesin tersedia dalam platform GEE, tetapi kebolehpercayaan pengkelas ini dalam pemetaan kolam akuakultur di kawasan tropika tidak dikaji dengan baik. Oleh itu, kajian ini bertujuan untuk menilai prestasi pengelas pembelajaran mesin terbina dalam seperti *Random Forest* (RF), *Support Vector Machine* (SVM), dan *Classification and Regression Tree* (CART) untuk memetakan kolam akuakultur di Sungai Udang, Pulau Pinang. Kemudian, imej Landsat bersama pengelas terbaik digunakan untuk mengesan pengembangan kolam akuakultur dari tahun 1989 hingga 2020. Imej Landsat dipilih kerana ketersediaan data yang lebih lama berbanding data satelit lain yang terdapat dalam platform GEE. Berdasarkan ketepatan keseluruhan, ketepatan pengeluar, ketepatan pengguna dan pekali kappa, ketiga-tiga pengelas pembelajaran mesin mampu menghasilkan ketepatan lebih daripada 80%. Walau bagaimanapun, ujian McNemar menunjukkan nilai yang signifikan antara SVM dan CART untuk tahun 2008 dan 2014, menunjukkan prestasi mereka tidak sama berkesan. Secara keseluruhan, RF ialah pengelas pembelajaran mesin terbaik untuk pemetaan kolam akuakultur di Sungai Udang, Pulau Pinang, dengan ketepatan lebih 90% untuk setiap tahun yang dinilai. Hasil kajian menunjukkan bahawa kawasan kolam Sungai Udang

telah diperluaskan daripada 150.52 ha pada 2004 kepada 311.59 ha pada 2020, terutamanya di bahagian barat Sungai Udang. Penemuan ini boleh bertindak sebagai rujukan kepada pihak berkuasa tempatan dan pengurus akuakultur untuk menguruskan kolam akuakultur mereka dengan berkesan.

COMPARISON OF GOOGLE EARTH ENGINE-BASED MACHINE LEARNING CLASSIFIERS FOR MAPPING AQUACULTURE PONDS IN SUNGAI UDANG, PENANG

ABSTRACT

Remote sensing is one of the most feasible ways to monitor and manage aquaculture activities. Aquaculture pond mapping can be conducted more easily and quickly due to rapid advancements in the cloud computing system, i.e., Google Earth Engine (GEE). Several machine learning classifiers are available in the GEE platform, but the reliability of these classifiers to map aquaculture ponds in tropical regions is still not well-studied. Thus, this study aims to evaluate the performance of different built-in machine learning classifiers such as Random Forest (RF), Support Vector Machine (SVM), and Classification and Regression Trees (CART) to map aquaculture ponds over Sungai Udang, Penang. Then, the Landsat images together with the best classifier were used to detect the expansion of aquaculture ponds from 1989 to 2020. Landsat images were selected due to the longer data availability compared to other satellite data that available in the GEE platform. Based on the overall accuracy, producer accuracy, user accuracy and kappa coefficient, all three machine learning classifiers are able to produce more than 80% accuracy. However, the McNemar test showed significant values between SVM and CART for the years 2008 and 2014, indicating their performance is not equally effective. Overall, RF is the best machine learning classifier for aquaculture pond mapping over Sungai Udang, Penang, with more than 90% accuracy for each evaluated year. The results show that the pond area of Sungai Udang was expanded from 150.52 ha in 2004 to 311.59 ha in 2020, mainly

in the western part of Sungai Udang. These findings can act as a reference for local authorities and aquaculture managers to manage their aquaculture ponds effectively.

CHAPTER 1

INTRODUCTION

1.1 Motivation and Background

The world's fastest-growing industry for producing food is aquaculture, which will soon dominate the food supply for people. Due to the world's population intensification, the ongoing demand of fish and foreign trade have all contributed to aquaculture's extraordinary rise in terms of production volume and value during the past few decades (Ottinger et al., 2016).

The output of aquaculture worldwide has increased by 609% between 1990 and 2020, growing at a yearly annual rate of 6.7%. Figure 1.1, which displays data from 1990 to 2020, demonstrates the rapid growth of aquaculture. With 91.6% of the world's aquatic animals and algae produced in Asia in 2020, the region has overwhelmingly dominated aquaculture for decades. With a cumulative growth of 22%, or roughly 19 billion kg, compared to 2020, aquaculture development is predicted to raise to 106 billion kg (53%) in 2030, as shown in Figure 1.2 (FAO, 2022).

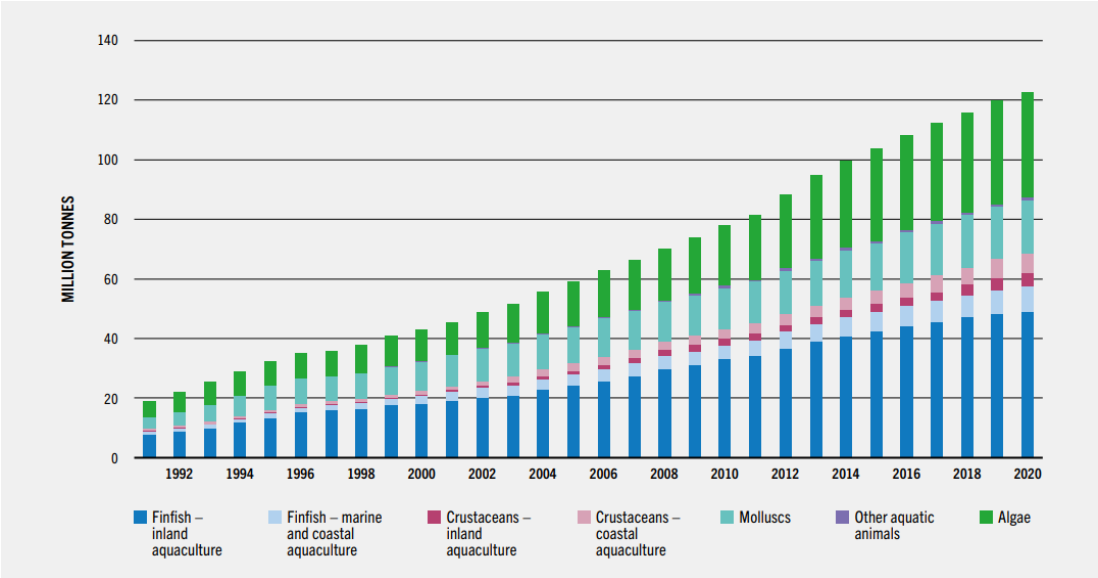

Figure 1.1 Global aquaculture production from 1990 to 2020 (FAO, 2022).

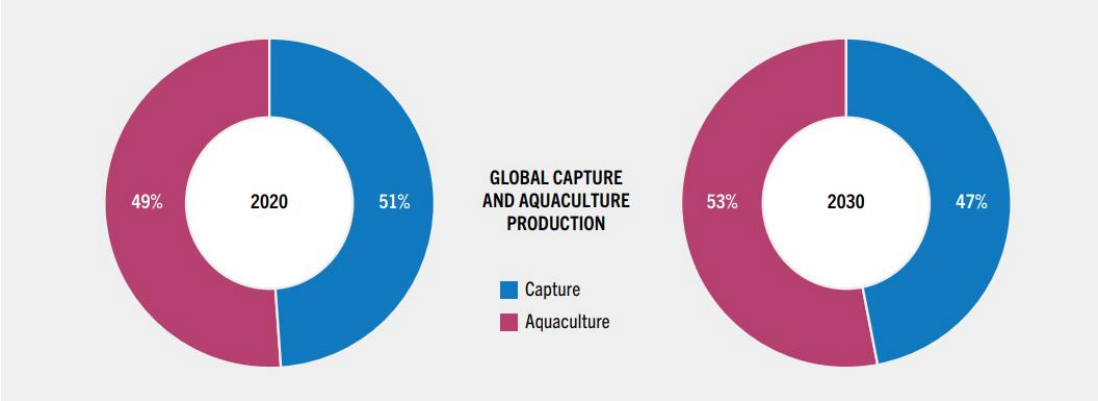

Figure 1.2 Global capture and aquaculture production growth prediction (FAO, 2022).

According to the Department of Fisheries Malaysia, aquaculture industry has produced 24 % of total fisheries output in 2021. A rough estimate of 417,187 million metric tonnes and RM3.43 billion in revenue are produced by Malaysia's aquaculture industry (DOF, 2021). This shows the aquaculture industry serves a crucial function in benefiting economically to Malaysia and also providing food security. Aquaculture is also beneficial in terms of socioeconomic. The sector reduces poverty by providing jobs. The environment has been harmed as a result of the sector's expansion to satisfy current demand, resulting in decreased availability of land, water contamination, eutrophication, hazardous chemicals, and food chain concerns (Troell et al., 2017). Currently, climate change has affected aquaculture in terms of production (FAO, 2022). The aquaculture industry's fast worldwide development has transformed enormous regions of important coastal and inland ecosystems, resulting in the depletion of goods and solutions provided by systems of natural features (Pattanaik & Prasad, 2011). Despite the fact that aquaculture management has become more important, an uncoordinated aquaculture development has resulted in severe

environmental destruction over the past years. This is due to poor environmental legislation and a lack of effective planning and supervision initiatives at the national and worldwide policy degree (Smith et al., 2010).

Various international organisations, including the FAO, World Fish, the United Nations, as well as non-governmental organisations (NGOs) and local governments, have acknowledged the growing relevance of aquaculture monitoring and management. Traditional ground surveys can be labour-intensive and expensive to carry out. Thus, remote sensing has the ability to help aquaculture monitoring. For instance, aquaculture mapping of historical and current data for the best site selection. In addition, aquaculture water quality monitoring and stock development have the potential to benefit aquaculture management. Satellite data are incredibly helpful for monitoring and assessing the aquaculture ponds, which act as important information for appropriate environmental conservation and natural resource control. Aerial assessment and mapping of aquaculture activities have taken a lot of time and effort in various study regions across the world. Remote sensing has a huge potential in aquaculture management because it provides critical supplemental information for national and international decision-makers and policymakers. Remote sensing helps us to understand what has been displaced as a result of the construction or intensification of aquaculture areas and how these changes have impacted the ecosystem (Ottinger et al., 2016).

The most practical way to get information about land use land cover (LULC) is to classify remote sensing data (Shalaby & Tateishi, 2007). In order to classify an image, pixels must be categorised based on a number of elements, including spectral signatures, indices, and contextual data. Maximum Likelihood Classification (MLC) is one of the popular parametric classifiers because to their superior outcomes (Yu et

al., 2014). However, parametric classifiers imply a standard data distribution, which does not relate for data from the actual world. Machine learning classifiers have become powerful classifiers in the last decade, and they have been extensively employed for LULC classification owing to their greater accuracies and performances than MLC (Ghimire et al., 2012). The non-parametric classifiers make no presumptions about how the data are distributed. Using remotely sensed images, non-parametric machine learning classifiers like Random Forest (RF), Support Vector Machine (SVM), and Classification and Regression Trees (CART) have been shown to provide results for LULC classification that are very accurate. (Foody & Mathur, 2004; Nery et al., 2016).

Analysing the reliability of different machine learning classifiers on multiple satellite images at a sizeable scale requires the use of powerful machines that can handle massive quantity of data and perform complex computations in a short amount of time. Only a select few people have access to such powerful computer settings. Additionally, it reduces the requirement for larger-scale revisions to LULC maps. Recently, a few cloud-based platforms are now available to help with large data and computing issues. These cloud-based platforms prepares the necessary tools for handling large amounts of data. These platform prepares a collection of images for classifying LULC. The Google Earth Engine (GEE), which has been employed in a few classification studies for particular LULC applications relevant to agricultural and urban areas at both the local and worldwide sizes, has been shown to be a successful platform (Dong et al., 2016). GEE allows several users to work alongside on the same pre-processed data collection, allowing them to reuse or test the ideas. The GEE platform provides built-in machine learning classifiers to classify images according to the user's needs.

1.2 Problem Statement

To produce a highly accurate aquaculture ponds map, it is essential to classify satellite images to create a LULC map using the best machine learning classifier. An assessment of the performance of various GEE built-in machine learning classifiers for aquaculture pond mapping in tropical regions is still lacking, which is crucial to obtaining the most accurate aquaculture pond map. Hence, this research aims to identify the best classifier that available in the GEE platform.

Analysis of the expansion of aquaculture ponds is crucial to understanding the development of aquaculture ponds. However, the aquaculture pond expansion studies are still lacking due to the lack of an appropriate framework to produce accurate aquaculture maps quickly. Hence, this study would like to answer the following questions:

1. What is the best performing GEE's built-in machine learning classifiers for aquaculture pond mapping over Sungai Udang, Penang?
2. What is the area expansion of aquaculture ponds in Sungai Udang, Penang over the past 32 years?

1.3 Objectives

The overall goal of the study is to produce aquaculture maps over Sungai Udang, Penang using the satellite data and classifier within the Google Earth Engine (GEE) platform. The specific objectives include:

1. To compare the performance of GEE's built-in machine learning classifiers in aquaculture pond mapping over Sungai Udang, Penang.

2. To detect the expansion of aquaculture ponds in Sungai Udang, Penang over a period of 32 years (1989-2020).

1.4 Scope of Research

This research compared the performance of GEE's built-in machine learning classifiers in producing aquaculture pond maps of Sungai Udang, Penang. Landsat data with 30 m of spatial resolution was utilised in this research due to the data availability since the past few decades. RF, CART, and SVM machine learning classifiers were utilised to classify the aquaculture pond maps. Standardisation of the method, data and parameters used was required to compare the three classifiers.

In Malaysia, pond systems are used in freshwater and brackish water environments. The pond system is an old one, yet it's still the most common way to create aquaculture products. Sungai Udang, Penang consists of only brackish inland aquaculture ponds. Thus, this study only focuses on extracting information from brackish inland aquaculture ponds in Sungai Udang, Penang.

1.5 Thesis Outline

This chapter discusses the introduction to the benefits of aquaculture, the expansion of aquaculture on a global and local scale, and the problems associated with it. This chapter also discusses some brief details regarding aquaculture pond mapping. The objectives and scope of research are explained in sections 1.3 and 1.4, respectively. The literature review of aquaculture, aquaculture in Malaysia, aquaculture ponds, remote sensing applications in aquaculture pond mapping, cloud computing and previous studies regarding aquaculture pond mapping using Google Earth Engine are discussed in Chapter 2. Chapter 3 discusses the data and the techniques utilised to conduct this study. This chapter also includes the study area and

the research's framework. The results obtained from the technique are explained and discussed in Chapter 4. This chapter includes the figures and tables obtained to visualise and summarise the results. Finally, the whole research is concluded in Chapter 5 and several recommendations are given for improvements for prospect ventures.

CHAPTER 2

LITERATURE REVIEW

2.1 Aquaculture

The increase of aquaculture production has been mostly attributed to the high food demand since 1980s. In all, 182 million metric tonnes of aquatic food are expected to be available for human consumption by 2030, a rise of 24 million metric tonnes from 2020. The majority of the intensification in worldwide fisheries will come from the aquaculture sector, whose production is anticipated to emerge 100 million metric tonnes for maybe the initial period around 2027. In 2030, it is anticipated that aquaculture production would increase to 106 million tonnes, a 22% total increase, or almost 19 million metric tonnes, from 2020. While capture fisheries' output has mostly stalled since the end of the 1990s, production from aquaculture has expanded dramatically. A combination of growing demand brought on by urbanisation, rising incomes, and the intensification of fisheries and aquaculture output will be the main driving reasons behind the rise in aquatic food consumption globally. Aquaculture has increased food security and enhanced nutrition in a large number of developing countries, specifically in Asia. The sector's benefits on lives and workforce are predicted to expand. Aquaculture must be sustainable, and inclusive, thus it is important to scale up revolutionary changes to management, investment, and policy (FAO, 2022).

Climate change has a significant impact on aquaculture industry. These dangers may amplify if there is a poor governance, which leads to habitat loss and environmental deterioration. Many evidences have proven the impacts of climate change on aquatic ecosystems. Therefore, a closer relationship between adaptation techniques and aquaculture management efforts is necessary, as well as a robust

evaluation of the climate change impact on the aquaculture sector. Fish farmers are already coping with the consequences of climate change in a variety of ways, including by diversifying their revenue sources, changing the way they catch and raise fish, and adjusting to environmental changes. Institutions and managerial structures must shift more swiftly, though. This necessitates the development of transformative adaptation plans at the national and regional level. These strategies must allow for autonomous adaptation in the medium and long term in order to encourage the transition of fisheries and aquaculture to a society that can adapt to climate change and to help achieve common aims of ending poverty and ensuring food security. (FAO, 2022).

Spatial management techniques may be used to plan, modify, and minimise the risks associated with current and future climatic conditions in the fisheries and aquaculture industry. Without appropriate spatial management and planning, significant disease breakout patterns and geographic organism's populations and habitats would alter as oceans warm and grow more acidic. Spatial planning and monitoring offer a solution-centred approach to better comprehend and forecast how fisheries and aquaculture may be impacted by climate change alongside to provide awareness into regional differences so that suitable region-based adaption measures may be implemented. By using spatial technologies like satellite remote sensing, aerial surveys, geographic information systems (GIS) and global positioning systems (GPS) excellent spatial planning and proper monitoring strategies at the area and farm management levels can lower risks posed by climate change and foster adaptation. Furthermore, it's critical to carefully design the temporal and spatial level of components for fish farming to make sure they are in accordance with the appropriate strategies for climate change adaptation and mitigation (FAO, 2022).

Due to the usual Malaysian interest and requirements in the fisheries business, which until recently was mostly focused on marine catch, there is a lot of stress to intensify output from the marine sector. Despite Malaysia having a monopoly on marine capture fish, the trend will not continue to improve and may even worsen in certain years due to the fact that most fish are collected in coastal areas. Thus, government assistance encourages the growth of aquaculture industries (Kurniawan et al., 2021).

Freshwater fish, seaweed, and brackish water fish are instances of products from Malaysian aquaculture, as are ornamental fish and aquatic plants. Malaysian aquaculture has experienced a modest shift as a result of the adoption of both conventional and traditional practices. Ponds, cages, cockles, mussels, and a long line for growing seaweed are all components of the brackish water and marine environment production system. More than 24352 fish farmers and aqua culturists will be working in this industry in 2021. Up to 311284 metric tonnes of fish, worth more than RM 2.57 billion, were produced through brackish water aquaculture. More than RM 860 million worth of fish, or 105904 metric tonnes, were produced through freshwater aquaculture. The output of seaweed for aquaculture was 178896 metric tonnes, worth RM 58 million. A total of RM 534 million worth of ornamental fish were produced in 242 million pieces, while RM 21 million worth of aquatic plants were produced in 21 million bundles (DOF, 2021).

The pond system as shown in Figure 2.1 and Figure 2.2, is one of the oldest and most established, yet it continues to be the best and most often used method for producing aquaculture products. The area of ponds ranges from 100 to 100,000 m^2, with a typical depth of 1.2 m to 1.5 m, depending on aspects including output volume, site conditions, and species category (Ngo et al., 2017). Benefits of this system include

ease of use, low worker and energy requirements, while negatives include dependency on outside factors like the weather and fierce competition in the fish market (Ahmad et al., 2021).

Figure 2.1 An aquaculture pond in Sungai Udang.

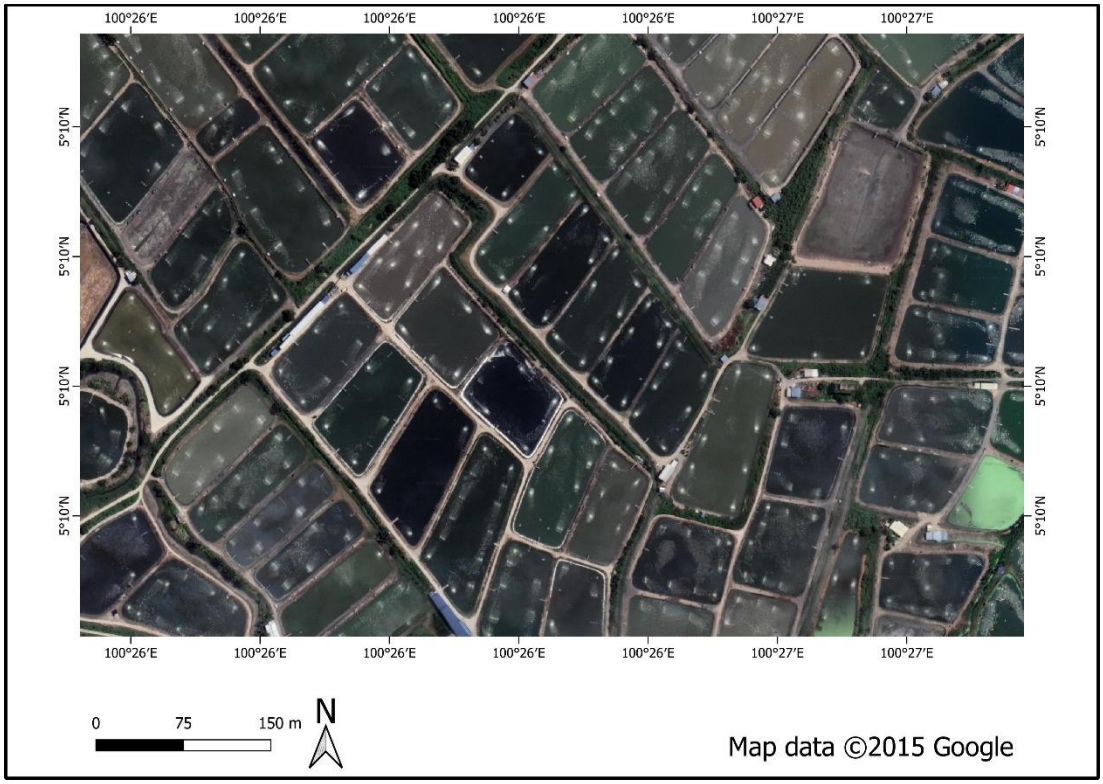

Figure 2.2 Aquaculture ponds top view taken from Google Earth Pro.

2.2 Remote Sensing

Remote sensing is the study of acquiring details about a distant object by exploiting reflected and emitted radiation. The majority of the time, sensor data is acquired using cameras, which provide an image. Airborne, land-based, and ground-based systems all provide remote sensing data (Barbosa et al., 2015). Additionally, remote sensing sensors utilised to acquire data passively or actively. Passive remote sensing collects data on the ground by using sunlight. The sensor will detect the signal from the ground. Active remote sensing does generate a signal, which is the information sent to the sensor by the object's reflected radiation (Joshi et al., 2016).

Multispectral remote sensing data generally consists of between three and ten bands, each of which is quantified by the wavelength of reflected energy. The visible bands (red, green and blue) and near-infrared (NIR) bands are some of them. Remote sensing data is available at a broad range of spatial resolutions, from low to high. Spatial resolution quantifies the degree of specificity seen by the human eye at the pixel level. The better the spatial resolution, the more precise the information that can be explored and retrieved (Qu et al., 2017; Yokoya et al., 2017).

2.2.1 Remote Sensing in Aquaculture Mapping

Remote sensing is a less expensive alternative to local governments' intensive field surveys and records, and it provides an immediate overview of broad sections of the world surface. Furthermore, relying on the sensor's revisit time, satellite data can be obtained on a frequent basis and can cover even the most remote places in developing nations, where land access may be difficult and restricted. As a result, remote sensing is an excellent method for spatially assessing aquaculture regions at various scales. Landsat and Sentinel are space-borne multispectral optical sensors with greater than 3 spectral bands that are highly valuable in monitoring and mapping aquaculture. Water has a low reflectance compared to other surface features because it absorbs shortwave and mid-infrared wavelengths of the electromagnetic spectrum. As a result, multispectral data is extremely effective for distinguishing between water and non-water surfaces. Many satellite data are costly or challenging to acquire, however there are more and more free and open access data available to help with upcoming aquaculture usage (Ottinger et al., 2016).

2.2.2 Techniques Used to Map Aquaculture Ponds

Long-term and large-area aquaculture pond mapping with Landsat imagery has proven to be more beneficial and cost-effective. Many studies have been accomplished utilising Landsat imagery at various spatial scales, which are discussed in this section. Pattanaik et al. (2008) studied aquaculture dynamics in Kolleru Lake using Landsat for the year 1977 and 2000. The study used Maximum Likelihood pixel-based supervised classification to produce classified maps for both years and found the aquaculture area increased to 158.5 km^2. However, machine learning classifiers have been reviewed and found to generate better accuracy compared to parametric classifiers.

Yao (2013) used a support vector machine classifier for land use mapping that included aquaculture ponds as one of the classes and calculated the area of aquaculture pond conversion for the years 1990-2000 and 2000-2010. A Similar LULC study conducted by Liu et al. (2020) used hybrid classification that includes a CART classifier and visual modification to produce a LULC map. The map includes aquaculture ponds class from 1979 to 2014 and the area of the aquaculture pond expansion. Faruque et al. (2022) used SVM, RF and SmileCart algorithms to classify LULC maps, which include aquaculture ponds as one of the classes, and found SVM provided high accuracy, which was then utilised for the classification of the years 1990 to 2020 with more than 90% accuracy. The area for aquaculture pond expansion and reduction was identified. These studies show the capabilities of machine learning classifiers to produce high-accuracy aquaculture pond maps using Landsat images.

A more advanced technique in aquaculture pond mapping by using object-based analysis conducted by Ren et al. (2019) utilised an object-based as well as integrated updating classification approach that included image segmentation (scale,

shape and compactness), rule building and manual editing to create maps of China's coastal aquaculture ponds between 1984 and 2015. According to the findings, China's coastal aquaculture ponds already span an area of 10463 km^2. However, the methodology requires many manual interventions by remote sensing experts. This process is time and processing intensive. Thus, object-based analysis is tedious and difficult to conduct.

Spectral indices were used by several studies to obtain a more accurate aquaculture pond map. For example, Sayah et al. (2020) utilised NDWI obtained from Landsat images to track changes in pond numbers with time in the Claise watershed. The accuracy of pond counts on the result map was 85.74%, and the spatial allocation of ponds was 75%, according to an analysis of aerial photography. Rani et al. (2021) evaluated spectral indices for separation of active aquaculture ponds from inactive aquaculture ponds utilising Landsat 8 images in the coastal of Guntur district. This study found NDWI, Water Ratio Index (WRI) and combinations of these indices performed better than the Maximum Likelihood classifier, NDVI and MNDWI. Thus, including suitable spectral indices can increase the accuracy of aquaculture pond maps.

Diniz et al. (2021) used a Deep Learning approach for long-term aquaculture pond mapping. In this study, Convolutional Neural Networks (CNN) and U-Net classifiers were used to obtain the Brazilian Coastal Zone aquaculture pond expansion area from the years 1985 to 2019. In this study, the processing time using the computer engine required 192 hours. The author reported that time constraints were significant because cloud services do not allow enormous processing power free of charge. Thus, this approach requires time and additional cost to be conducted.

High-spatial resolution image has been utilised to map aquaculture in some studies. Chen et al. (2006) used SPOT 5 images for land cover mapping, which

included aquaculture ponds class. A stacked unsupervised classification approach was utilised in this work to determine the area of aquaculture. Virdis (2014) used SPOT 5 and Worldview 1 to map the aquaculture ponds with the RGT segmentation algorithm and ISOSEG classifier, and both images produced overall accuracy of more than 80%. However, the author reported the appearance of salt and pepper in the images during the segmentation process. Thus, high-resolution images are not suitable for pixel-based classification, as the salt and pepper issue may reduce the accuracy of the map. A study by Nguyen-Van-Anh et al. (2021) compared VNREDSat 1 (2.5 m) and Sentinel 2 (10 m) for LULC mapping, which includes aquaculture pond class. This study concluded that by using same the methodology for both images, aquaculture ponds were effectively classified with Sentinel 2 because of their similarities.

Sentinel satellite provides both optical (Sentinel 2) and microwave images (Sentinel 1). A study by Ssekyanzi et al. (2021) used Sentinel 2 images to map the inland aquaculture pond using several water index methods and the found Automated Water Extraction Index without shadow pixels (AWEIsh) and Automated Water Extraction Index with urban background (AWEInsh) had the highest overall accuracy. A study conducted by Haris et al. (2021) compared Sentinel 1 and Sentinel 2 data for aquaculture ponds mapping using a machine learning classifiers and found Sentinel 1 imagery produced better overall accuracy compared to Sentinel 2 imagery. However, the methodology used for Sentinel 1 imagery was tedious and time-consuming compared to Sentinel 2 imagery. Sentinel images are only available from 2014, which is not suitable for longer temporal studies. Thus, Landsat imagery is the ideal option for a long-term aquaculture pond mapping study.

2.3 Cloud Computing

Earth observation data, also known as remote sensing data, have increased dramatically in number, variability, and complexity and are now referred to as remote sensing "Big Data." Remote sensing professionals use the phrase "Big Data" to describe data that is so large, fast-moving, diverse, and sophisticated that it exceeds the capability of storage and processing systems. Additionally, it can take a while to handle the enormous amounts of remote sensing "Big Data," which encompasses loading, storing, analysing, and evaluating data. To analyse enormous amounts of data, great processing power will be needed, which will be expensive. Cloud computing is one of the best answers to this problem, which efficiently virtualizes supercomputers for users while processing distant sensing "Big Data" on extremely powerful servers. Cloud computing systems like Amazon EC2, Google Earth Engine as well as Microsoft Azure may all be utilised to manage "Big Data" from remote sensing (Ma et al., 2015).

2.3.1 Google Earth Engine

In comparison to Microsoft Azure and Amazon EC2, Google Earth Engine (GEE) is a free-to-use cloud computing platform. It offers unprocessed images, preprocessed, cloud-free images, and mosaicked images of up to 40 years' worth of petabyte-scale remotely sensed data from various satellites. Because the GEE platform is backed up by Google's computer infrastructure, which allows for parallel processing of geographical data, computational duration may be shortened. Users' API (JavaScript and Python) codes may be stored and shared in a Git repository on the Earth Engine servers, allowing for collaboration. A web-based Integrated Development Environment (IDE) for creating, coding, and running sophisticated JavaScript applications is GEE's code editor. The code editor in the GEE is very convenient and

includes a wide range of algorithms. Users do not need extra software for operations like image processing, image collection, geometry-feature reduction, and machine learning classifiers, to name just a few. The GEE Explorer is a web application that enables users to view and perform basic analytics on data libraries (Gorelick et al., 2017; Kumar & Mutanga, 2018; Tamiminia et al., 2020). GEE open-source cloud-based platform possesses the ability to provide a huge amount of geospatial data and easy processing and analysis without having to battle the many computationally related difficulties.

GEE has already been used in a number of study areas, as shown in Figure 2.3 because of its capacities. Gong et al. (2013), produced a 30 m worldwide land cover map using the GEE cloud computing platform. The analysis can be completed on a single platform. Midekisa et al. (2017) throughout a 15-year period, utilised GEE to generate yearly maps for the entire Africa. Hansen (2013) utilised GEE to collect 12-years of satellite images to track the change in the global forest cover at 30 m resolution and show global forest loss and gain.

Figure 2.3 Categorization of GEE application (Tamiminia et al., 2020).

2.4 Machine Learning Classifiers

To extract LULC information such as aquaculture ponds from remotely sensed data, image classification plays an important role. Image classification consists of allocating pixels into classes according to their spectral properties, indices, contextual information, and many more. Many classification approaches are available, of which Maximum Likelihood Classification (MLC) is the commonly used parametric classifier because it produces a good classification outcome (Yu et al., 2014). However, parametric classifiers imply distribution of data is normal, and usually, data does not adhere to this type of distribution, while non-parametric classifiers do not have any implication regarding the distribution of data. Machine learning classifiers have appeared as dominant classifiers recently, and they have been commonly employed for LULC classification due to their superior accuracy and effectiveness (Ghimire et al., 2012). Non-parametric machine learning classifiers, for instance,

Random Forest (RF), Support Vector Machine (SVM) and Classification and Regression Tree (CART) have been found to produce very accurate LULC classification outputs with remotely sensed images (Foody & Mathur, 2004; Nery et al., 2016). Thus, non-parametric machine learning classifiers such as SVM, CART and RF would be suitable to be applied to and compared for aquaculture pond mapping.

2.4.1 Classification and Regression Tree

Breiman et al. (1984) created CART, one of the simplest binary classifiers in the conceptual architecture of centralised decision trees. Such approaches have the essential advantage that classification decisions may be considered as white-box systems in which the input-output connections can be easily understood and examined (Tso & Mather, 2009).

As illustrated in Figure 2.4, a sequence of nodes that are each divided into two branches connect the input and output of the CART algorithms, eventually leading to leaf nodes that match the labels on classification trees and continuous variables in regression trees. Until a threshold condition is met, nodes are divided repeatedly. The Gini Impurity Index is used by CART to determine which input variables will result in the optimal split at each node (Tso & Mather, 2009). With input features ordered in a linear manner, the separation can be either univariate, with decision boundaries parallel to the input feature axis, or multivariate, with additional flexibility for each class border (Tsoi & Pearson, 1991).

When the training data is very well-fit, CART has a tendency to overfit the tree. Trimming the tree such that it can survive incoming non-training data is the solution to this issue. Cross-validation pruning is used by CART to remove branches whose removal has no impact on the outputs above a particular threshold (Lawrence & Wright, 2001). This may result in a loss of understanding and a drop in accuracy for

training data categorization, but it also increases the accuracy for unknown data (Pal & Mather, 2003).

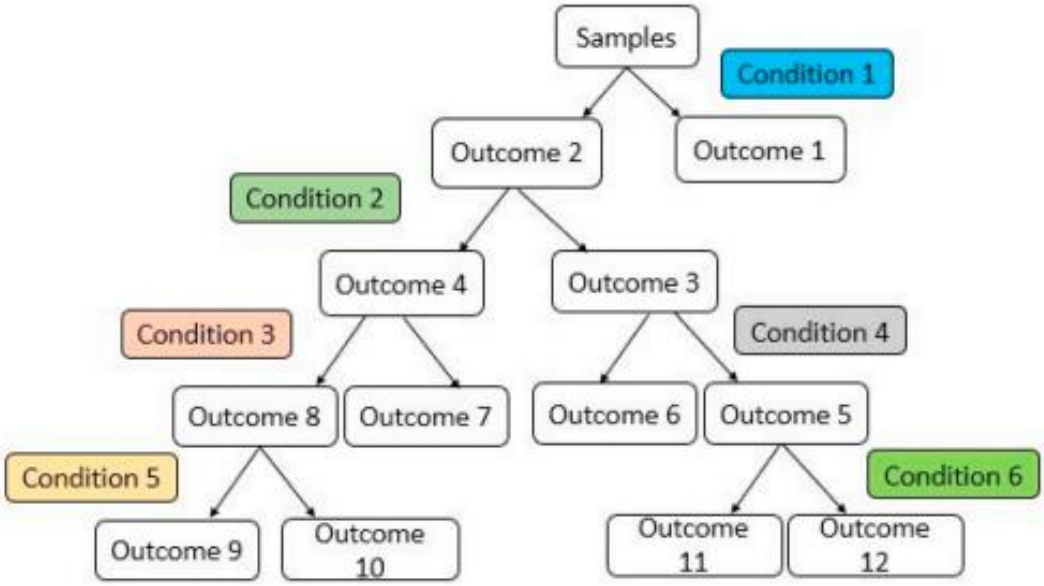

Figure 2.4 CART structure (Shaharum et al., 2020).

2.4.2 Random Forest

Tumer & Ghosh (1996) demonstrated that integrating the output of many classifiers to predict a finding result in extremely high classification accuracy. The RF ensemble classifier, which combines the output of several decision trees to select the label for new input data based on the best vote, is built on the previous statement, as shown in Figure 2.5. When building a single tree by randomly replacing a subset of training samples, RF uses the bagging approach, in which data is collected from the whole training set for each tree. This can result in some trees getting the same samples chosen while other trees get none at all (Breiman, 1996). The performance of the classifier is assessed internally using the non-training examples, which also provide an unbiased assessment of the classification error. Each node's proper split for building a tree is determined by RF using a random choice of variables from training samples.

This lessens the association between individual trees while weakening each one's power, which lowers classification error (Breiman, 2001). The Gini Index, a gauge of contamination in a node, is used by RF to estimate the ideal split. Entropy is reduced and information gain is increased after the split because of the manner the split is performed. However, the best split selection measure has less of an impact on the effectiveness of tree-based classifiers than the choice of pruning processed (Pal & Mather, 2003). Due of its ability to develop trees without the need of trimming techniques, RF is immune to these impacts (Pal & Mather, 2005).

A collection of trees takes into consideration all the characteristics that are randomly chosen from the training samples, but a single tree may not obtain the relevance from all input variables and may prefer some features during classification. RF assists in identifying the relative importance of numerous parameters derived from a satellite image's bands in the context of remote sensing. Each input variable is evaluated by RF by removing one and holding the rest constant from a number of input variables chosen at random. Based on out-of-bag error and a decline in the Gini Index, it determines accuracy (Ghosh et al., 2014). Additionally, RF determines the separation between two samples based on how frequently they arrive at the exact terminal node. This proximity analysis makes RF noise-insensitive and assists in the identification of incorrectly labelled training samples (Rodriguez-Galiano et al., 2012).

RF has become more common owing to its resilience to noise and outliers. Furthermore, RF exceeds other ensemble methods employed by other classifiers, such as bagging and boosting (Gislason et al., 2006). Even when utilised in a wide range of applications, such as urban landscape classification (Ghosh et al., 2014). Furthermore, LULC classification using SAR data (Waske & Braun, 2009). RF has been shown to be effective.

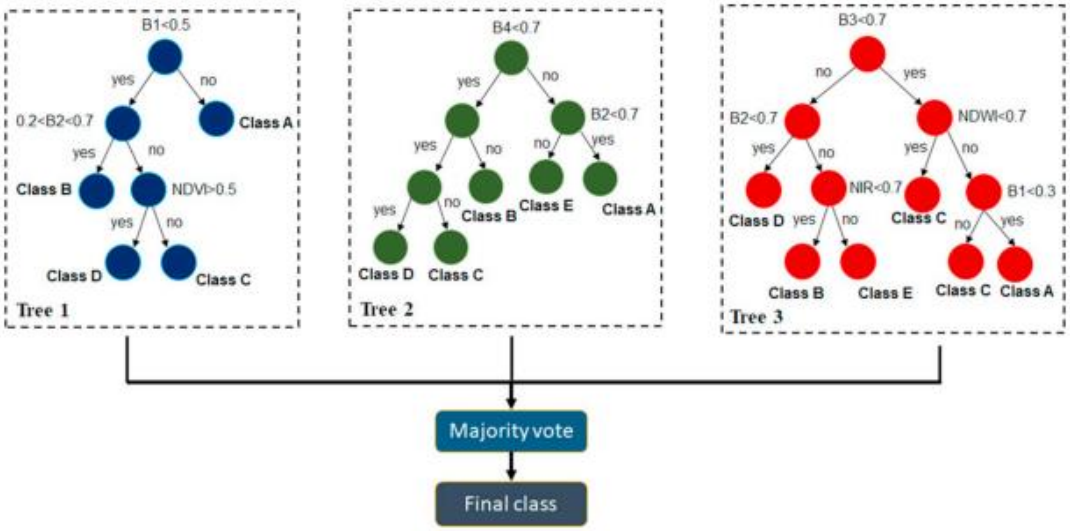

Figure 2.5 RF structure (Shaharum et al., 2020).

2.4.3 Support and Vector Machine

One of the machine learning classifiers used most commonly in the remote sensing sector is SVM. SVM became popular because of its ability to provide very accurate classification results with little training data, which is often a challenge in LULC classification applications (Mantero et al., 2005). SVM is a linear binary classifier that operates under the premise that training samples that are more closely associated with the boundaries of the class discriminate the class more effectively than other training samples. SVM concentrates on choosing the ideal hyperplane for classifying the input training data, as seen in Figure 2.6. To build support vectors, which are subsequently utilised in the real training, samples near a class's boundaries and the hyperplane are used. This is rarely the situation, though. A relief is provided for classes with non-linear connections in the form of the slack variable, which allows a few incorrect pixels to exist inside a class border while still reaching a hyperplane (Cortes & Vapnik, 1995).

SVM is a non-parametric machine learning classifier that is often employed since it generates precise results with a few examples of training data while effectively

applying to new input data. It works well with elevated data as more high-resolution, multi-spectral data becomes available, which is a big advantage in the area of remote sensing (Srivastava et al., 2012).

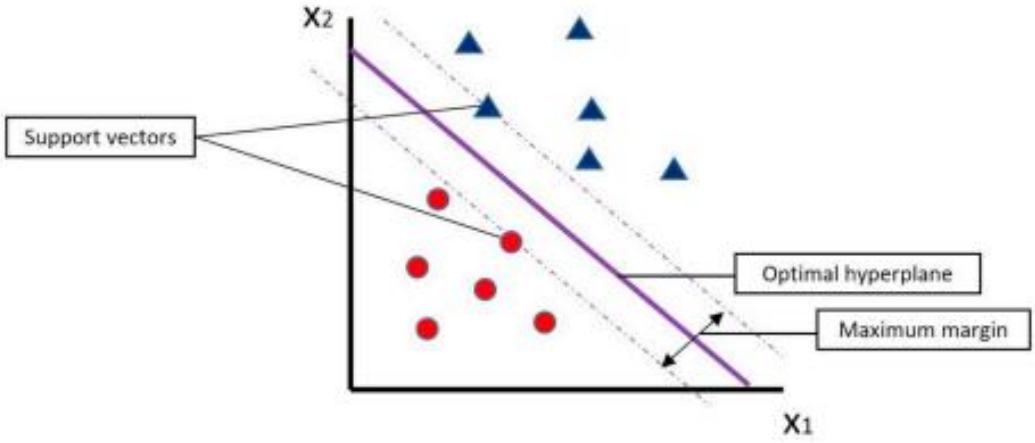

Figure 2.6 Optimal hyperplane finding in SVM (Shaharum et al., 2020).

2.4.4 Comparison of Machine Learning Classifiers

Numerous studies have examined the effectiveness of different machine learning classifiers for classification, such as LULC classification. Accuracy assessment is the best strategy for an analysis of classification performance (Lu & Weng, 2007). Some researchers have conducted studies to compare machine learning classifiers using Landsat, which is the imagery used in this study. For example, Shaharum et al. (2020) compared RF, SVM and CART classifiers to generate land cover maps using Landsat data and found all 3 classifiers produced good results, generating overall accuracy of 80.08% (CART), 93.16% (SVM) and 86.50% (RF). Furthermore, RF and SVM classifiers showed non-significant in p-values when compared, which shows these classifiers produce high accuracy maps. Ghayour et al. (2021) evaluated the performance of SVM, Artificial Neural Network (ANN), Maximum Likelihood Classification (MLC), Minimum Distance (MD) and