Machine Learning Based Approaches for Classification of Oil Spills and Microplastics in Marine Environments

by

© Yifu Chen

A thesis submitted to the

School of Graduate Studies

In partial fulfillment of the requirements for the degree of

Master of Engineering

Civil Engineering

Faculty of Engineering and Applied Science

Memorial University of Newfoundland

May 2021

St. John's Newfoundland and Labrador

ABSTRACT

Environmental modelling is an important approach of environmental engineering and management since it helps gain better understanding of environmental problems and impacts and facilitate environmental decision-making processes. However, because of the intricate conditions enormous data, diverse uncertainties, and various standards and requirements, environmental modeling is usually sophisticated and challenging. This study aimed to develop the novel modelling approaches by integrating machine learning (ML) into analyzing tabular and image datasets for environmental applications.

Firstly, a data-driven binary classification approach was developed to analyze oil fingerprinting. After comparing six different machine learning algorithms on five different biomarkers, random forest classifier was found as the most effective and accurate model to distinguish weathered chemically dispersed and non-dispersed oil from the dataset of diamantanes. The developed model was approved to be capable of aiding oil fingerprinting under the studied conditions. It showed the good value of ML methods in environmental modeling especially for oil spill response research and practice.

Secondly, an integrated approach by combing the strengths of convolutional neural networks and improved deep convolutional generative adversarial networks was proposed to classify microplastics and oil-dispersant agglomerates (MODAs) with diverse weathering conditions. The f score and model accuracy suggested the robust prediction from the trained model on the dataset of MODAs with different weathering degrees. The results could provide a better understanding of microplastics' effects on oil fate and transport during a marine oil spill. The proposed approach also presented the high potential of facilitating image-related classification work in environmental fields.

This dissertation not only developed two new ML based modelling approaches for environmental applications in oil fingerprinting and oil/microplastics classification, but also demonstrated the high value of ML methods and deep neural networks in processing experimental data for supporting environmental engineering and management.

ACKNOWLEDGEMENTS

It is worth to note that this dissertation research was conducted despite the impact by the COVID-19 pandemic. First and foremost, I want to express my heartfelt thanks to my supervisor,

Dr. Bing Chen, for giving me with the invaluable chance to earn my master's degree, for his patient direction, support, and encouragement throughout my studies, and for his profound life and professional lessons. I am especially grateful for the support from the Northern Region Persistent Organic Pollution Control (NRPOP) Laboratory, the Faculty of Engineering and Applied Science and Memorial University, Natural Sciences and Engineering Research Council of Canada (NSERC) and its Collaborative Research and Training Experience (CREATE) program for the Network on Persistent, Emerging, and Organic Pollution in the Environment Program (PEOPLE), and Canadian Foundation for Innovation (CFI). Additionally, I would like to give acknowledgements to Dr. Baiyu Zhang for her kind advice, as well as the members of our research group, including Dr. Bo Liu, Dr. Xing Song, Dr. Zhiwen Zhu, Dr. Weiyun Lin, Dr. Xiaying Xin, Xudong Ye, Min Yang, and Qiao Kang for their helpful assistance and suggestions during my study.

Finally, I would like to express my gratitude to my parents and family for their unwavering love, support, and encouragement. Especially in the past two years of pandemic, thousands of calls witnessed how they cheered me up through the journey in my life when I am away from my loved ones from the distance of 11,184 km. Without their assistance and care, I won't make it to the end of my graduate study.

Over the past two years, with the love and support from all the above people, I am proud of my progress and growth both physically and mentally which I have learned from all the academic and daily life's happiness and hardship. Now, I can finally say that I have made it and much look forward to starting a new chapter of my career and life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS AND SYMBOLS

ABBREVIATIONS

AI	Artificial intelligence
ANN	Artificial neural networks
ATR	Attenuated total reflection
BigGANs	Big generative adversarial networks

BN	Batch normalization
BRT	Boosted regression trees
CDO	Weathered chemically dispersed oil
CNNs	Convolutional neural networks
DCGANS	Deep convolutional generative adversarial networks
cGANs	Conditional generative adversarial networks
DL	Deep learning
DOR	Dispersant-to-oil volumetric
DTC	Decision tree classifier
EDS	Energy Dispersive X-ray Spectroscopy
ESE	Environmental science and engineering
ESI-LC-MS	Electrospray ionization liquid chromatography-mass spectrometry
EVC	Ensemble vote classifier
FT-ICRMS	Ultrahigh-resolution Fourier transform ion cyclotron resonance mass spectrometry
FTIR	Fourier transform infrared
GANs	Generative adversarial networks
GC/MS	Gas chromatography and mass spectrometry
IR	Infrared spectroscopy
IRMS	Isotopic resolution mass spectrometry
KNN	K-nearest neighbor

LIME	Local interpretable model-agnostic explanations
LRC	Logistic regression classifier
InfoGANs	Information maximizing generative adversarial networks
ML	Machine learning
MPs	Microplastics
NRPOP	Northern region persistent organic pollution control laboratory
OLE	Ordinary least squares
PAHs	Polycyclic aromatic hydrocarbon
PCA	Principal component analysis
PPE	Personal protective equipment
ResNet	Residual networks
RF	Random forest
RFC	Random forest classifier
ROI	Region of interest
RSD	Relative standard deviation
SEM	Scanning electron microscopy
SHAP	Shapley additive explanations
SMOTE	Synthetic minority over-sampling technique
SN	Spectral normalization
SVC	Support vector classifier
VGG	Visual geometry group

WCO	Weathered crude oil
WMODAs	WMPs-oil-dispersant agglomerates
WMPs	Weathered microplastics
WOT	Without transfer learning
WT	With transfer learning
XGBoost	Extreme gradient boosting
YOLO	You only look once

SYMBOLS

f	Mapping function
x	Input variable
y	Output variable
PC_i	Principal component
X_d	Original feature d
a_d	Numerical coefficient of X_d.

D	The distance between the query and examples
x_i	The query
y_i	Examples.
D	Distance
K	The specific number of examples
N	The number of features
M	The number of regions in the feature space
R_m	A region suitable to m.
C_m	A constant appropriate to m
P	The probability of a label 1
e	The base of the natural logarithm
a, b	The model's parameters
y	The predicted label
Cn	Different classifiers
C	Regularization
TP	True positive
FP	False positive
FN	False negative
TN	True negative
F-score	Evaluation score
μ	Means batch mean
σ	Means batch standard deviation
Z	Means layers input.

$Z_N^{(i)}$	Means the normalized Z
ε	A constant used for numerical stability.
Z_O	Means output
γ	Allows to adjust the standard deviation
β	Allows to adjust bias
G	Generator
D	Discriminator

CHAPTER 1: INTRODUCTION

1.1 Background

Machine learning (ML) is a branch of artificial intelligence and computer science that focuses on using data and algorithms to replicate the way people learn, steadily increasing accuracy. Due to their powerful modeling capability, ML is a powerful primary tool for data scientists to analyze and interpret data (Jordan and Mitchell, 2015). In terms of prediction, there are two typical types of tasks that ML performs, classification and regression. The task of estimating a mapping function (f) from input variables (x) to discrete output variables (y) is known as classification predictive modeling. In contrast, the task of estimating a mapping function (f) from input variables (x) to a continuous output variable (y) is known as regression predictive modeling (Loh, 2011). The application of ML in data analysis-related fields has been booming these years due to the advancement of computer hardware, for example, from medical to financial and environmental research.

There have three main areas of ML, supervised learning, unsupervised learning, and reinforcement learning. Each of them specializes in different tasks. For example, supervised learning is suitable for classification and regression with the labeled dataset, while unsupervised learning is more commonly used with unlabelled datasets (Reinel et al., 2020). Unlike working with labeled datasets in supervised learning, reinforcement learning labels the sequences of dependent decisions through agent and reward systems, and it is mainly applied in the game and robotic fields (Vincent et al., 2018). Deep learning (DL), as a more advanced type of ML, has been implementing to handle complicated tasks, for example, synthetic data (Ziwei et al., 2020). Generative adversarial networks (GANs) have been one of DL's primary applications in generating synthetic data (e.g., images) since it was introduced by Ian et al. in 2014. GANs are composed of two neural networks, which are generator and discriminator. The generator takes random noise

from the gaussian distribution as input and passes the noise to several upsampling layers to the desired shape of images. Sequentially, the discriminator takes the fake output of the generator and real image dataset as input and passes them to downsampling layers and calculates the loss between fake output and real image dataset. The generator parameters will then be updated through backpropagation until the generated images are good enough to fool the discriminator (Wei et al., 2018). However, the initial version of GANs (a.k.a., vanilla GANs) is extremely difficult to train due to non-convergence, mode collapse, diminished gradient, overfitting from the unbalance between the generator and discriminator, etc (Hamed et al., 2019). Therefore, many modified versions of GANs have been proposed to optimize the problems, for example, deep convolutional GANs (DCGANs), conditional GANs (cGANs), information maximizing GANs (InfoGANs), Pix2Pix, Big GANs (BigGANs), etc. These optimized GANs are served in different purposes, from transferring portraits to animated profiles and enhancing old photo qualities (Tero et al., 2018; Andrey et al., 2020; Tomaso et al., 2020; Yuan et al., 2021).

Currently, there have many ML algorithms that can perform regression and classification on different types of datasets, e.g., sequential, text, audio, or images. Understanding how to choose the appropriate ML algorithms is essential, and a comparison of the advantages and disadvantages of different ML algorithms is needed to be provided. Moreover, some ML algorithms are black-box algorithms with low expandability and could cause low reliability for the end-users. Thus, how to improve the developed ML models' reliability is desired to be explored. In terms of GANs, as stated above, the training suffers from some problems. Even though some different structured GANs are proposed, further optimization still needs to be conducted when training happens in different domains.

ML has shown promise in tackling complex data patterns or formats due to its incredible fitting skills. As a result, over the last decade, ML, particularly DL, has experienced tremendous growth in various applications, including image categorization and machine translation. Researchers in the broad field of environmental science and engineering (ESE) have enthusiastically embraced ML in various applications. For example, Munish and Parveen compared empirical infiltration models with ML-based adaptive neuro-fuzzy inference system and random forest regression techniques on the soil infiltration rate dataset. They concluded that ML-based methods are the most appropriate technique for estimating the infiltration data (Munish and Parveen, 2019). Behrooz et al. (2020) utilized long short-term memory, a method based on deep neural networks, to model faults in the oxidation and nitrification process in wastewater treatment plants since the nonlinear dynamics and complex interactions of the variables in wastewater data. Their proposed model achieved a recall of over 92%, outperformed traditional methods, and enabled timely detection of collective faults. Hao et al. (2020) proposed an enhanced approach of generative adversarial networks to generate more environmental microorganisms (EM) images since EM analysis plays an essential role in environmental monitoring and protection. The generated images were further evaluated by average precision from ResNet50 and VGG16. Their results demonstrated that the proposed model could achieve remarkable performance in augmenting EM images with high quality and resolution, improving EM image classification precisions.

Despite the wide application of ML and DL in wastewater, soil, and other environmental fields, the applications in marine oil spill-related fields are still lacking. More examples and case studies of how ML and DL can help with marine oil spill-related problems are urgently needed because the ocean is the world's largest ecosystem.

As the primary energy source, fossil fuel supplied 84% of world energy in 2019, while in Canada, fossil fuel accounted for 88.5% of energy supply in 2018. Among all fossil fuels used in Canada, crude oil took the most significant portion of 47.1%, followed by natural gas of 32.2% (Carroll and Huijzer, 2018). However, during the production, transportation, and usage of fossil fuels, oil spills have increased. The major oil spills usually happen in the marine environment, e.g., the Deepwater Horizon spill, the largest marine oil spill in the U.S. in 2010, and many smaller scale of oil spills. Figure 1.1 presents the oil spills incidents from 2008 to 2020 in 7 -700 metric tons and more than 700 metric tons worldwide (Gracia et al., 2020). These oil spills can have some detrimental effects on the ecosystem and economy. The leaked oils are highly toxic, and ingestion or inhalation of these oils can cause damage to DNA, immune function and cardiac dysfunction, and mass mortality of eggs and larvae. The oil spills could negatively affect the tourism industry, port business, sea-based transportation, and fishery (Zhang et al., 2019). These damages are also valid in Newfoundland and Labrador (NL), the third-largest oil producer in Canada. NL generated 4.4% of Canada's petroleum from its Grand Banks offshore oil fields. In NL's history, the largest oil spill happened in November 2018. Husky Energy reported a spill of an estimated 250,000 liters of crude oil from their SeaRose platform, and it killed thousands of seabirds due to oil pollution (Higgins, 2011). It can be expected that oil spill accidents will continue to happen before totally switching to a non-fossil fuel-powered society.

Oil spills can happen anywhere and anytime, for example, onshore and offshore. In this thesis, offshore oil spills are mainly focused on. When an oil spill occurs in a marine environment, the impacts have wide-ranging implications as a long-lasting environmental disaster. Depending on the amount and preparties of the spilled oil and its location and ambient environment, the

impacts might vary significantly. For example, oils may affect animals' respiration, feeding and severely influence their habitat.

Moreover, the entire ecosystem might also alter remarkably and even permanently due to toxic chemical components of the spilled oil (Chang et al., 2014). Besides the ecological damage, oil spills can cause significant economic loss. The contamination of coastal areas can disrupt recreational activities in tourism and incur long-term economic damage when public perception of prolonged pollution remains long after the oil has gone (Palinkas, 2012). Damages are also usually observed in fishery and mariculture sectors and coastal community livelihoods (Gracia et al., 2020).

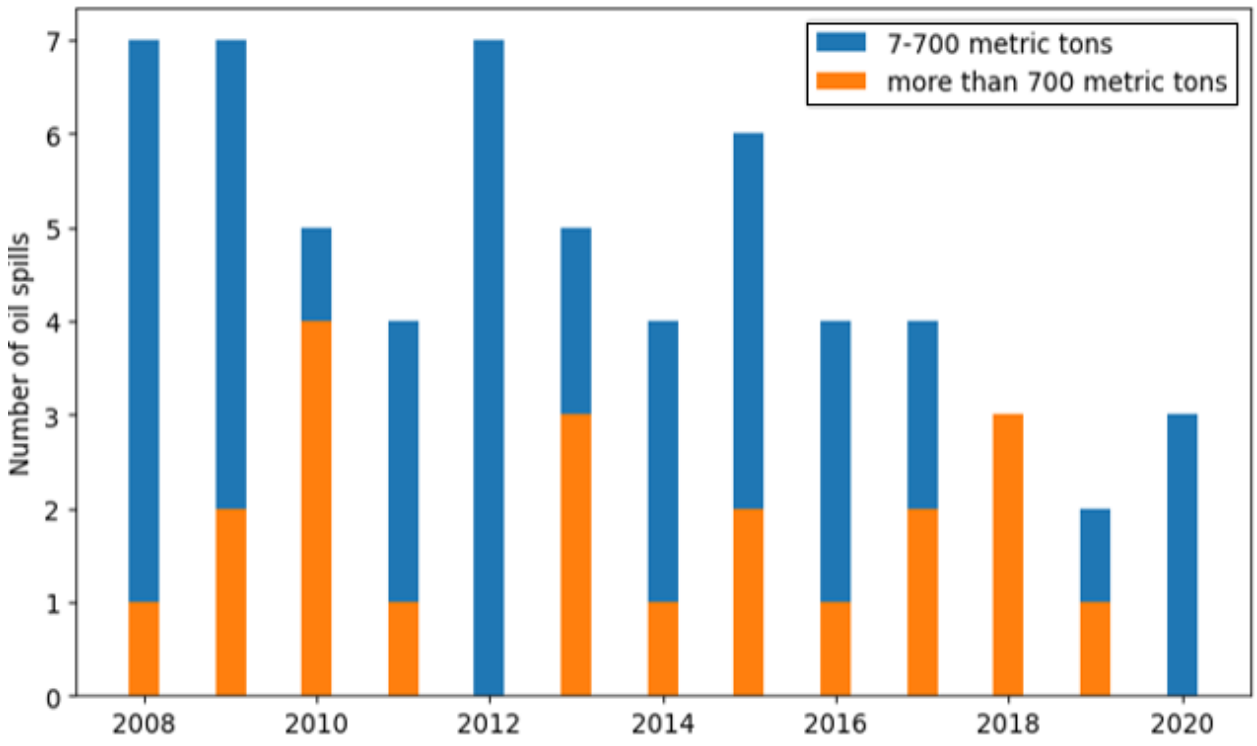


Figure 1.1 Global oil spills from 2008 to 2020

Tracking back the origin of oil spills is part of environmental forensic. Knowing where the spilled oil comes from helps investigate the cause of oil spills, for example, whether it is caused by accidents involving tankers, pipelines, drilling rigs, or the collision of ships. If there are no matching reports with oil spills, investigating the origin of spilled oil could minimize the oil leaking to the marine environment. Furthermore, it is essential to know who takes responsibility for oil spills when nobody makes an announcement (Wang and Stout, 2010). Currently, oil fingerprinting is mainly used in environmental forensics to trace oil spills' origin. Since the formation of hydrocarbons in oil and gas deposits is affected by many factors, such as temperature, reservoir tectonics, biodegradation, aquifer activity, etc. These factors are different in every reservoir, which allows the identification of hydrocarbons from evaluating changes in composition (Stout and Wang, 2016). Therefore, oil fingerprinting is based on geochemical analysis of hydrocarbon fluids composition, which could provide valuable and unique information to identify hydrocarbons' origin.

There are numerous counter measurements to marine oil spills, for example, booming, skimming, in-situ burning, and dispersion (Dave and Ghaly, 2011). This dissertation research mainly focused on the application of dispersants. An oil dispersant is a combination of emulsifiers and solvents that aids in the separation of oil by spraying on a surface oil slick to break down the oil into smaller droplets, allowing them to mix with the water more readily and enhance biodegradation by sea-living microbes (Brakstad et al., 2018). As stated earlier, the key for oil fingerprinting is hydrocarbons' composition because of their uniqueness. However, when dispersants are spilled in oil in a marine environment, the hydrocarbon composition might be changed after the application. This change poses a challenge to oil fingerprinting since it brings bias in the investigation in environmental forensic (Joo et al., 2013). Hence, applying different

protocols is necessary on dispersed oil and crude oil and distinguishing dispersed oil and crude oil is the first step before identifying their origins.

Besides the pollution of oil spills, the ocean also faces other negative impacts from emerging pollutants, and plastic is on the top of the pollutant list. Plastic has existed in human society for a long time since it was invented. Among all the plastic, 10% of plastic products will end up in the ocean (Magnier et al., 2019). Figure 1.2 illustrates the plastic in the global ocean and other individual oceans in 2013. In 2020, 300 million tonnes of plastic waste will be generated, and if 10% of them will end up in the ocean, there will be 30 million tonnes of plastic (Ostle et al., 2019). Compared to Fig 1.2, it increased 100 times from 2013 to 2020.

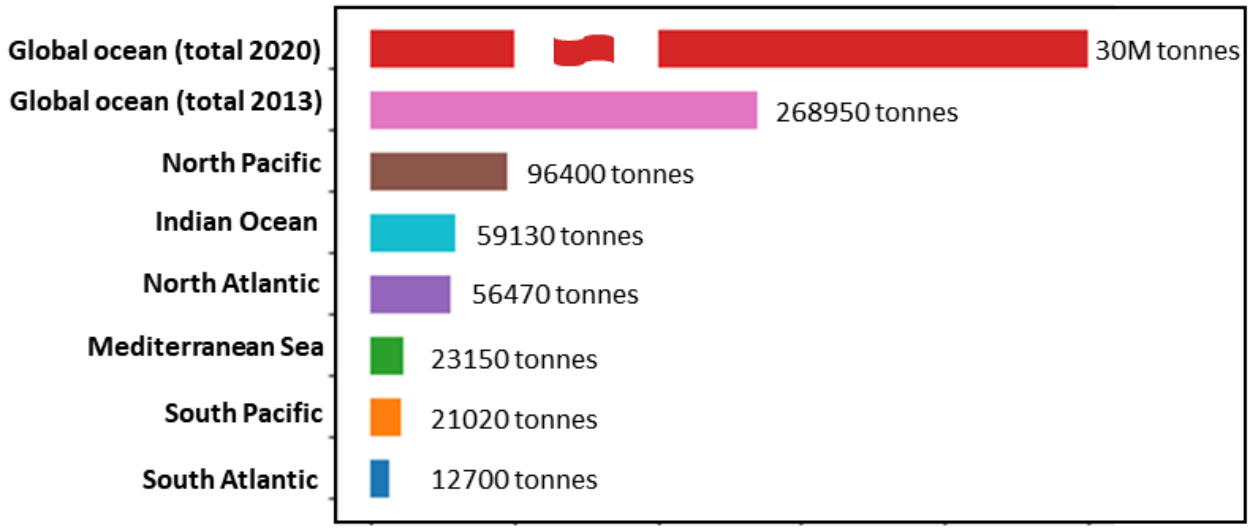


Figure 1.2 Surface plastic mass by different oceans in 2013 and global total (2013 VS. 2020)

As the main component of personal protective equipment (PPE), plastic usage spikes during the COVID19 pandemic. Under weathering process, including mechanical tension (e.g., wave motion), photooxidation, and biological degradation, these plastics will be fragmented into micro-level size (less than 5mm in length), which are considered as weathered microplastics

(WMPs) (Jahnke et al., 2017). In recent years, there has been a rising concern about negative impacts on the ocean ecosystem and economy caused by WMPs (Ronkay et al., 2021). WMPs are hard to degrade, and they can interact with another ocean pollutant, spilled oil (Liu et al., 2020). Spilled oil can be broken down into tiny droplets by dispersants and become much easier to degrade in the marine environment (Wu et al., 2021). However, when WMPs join the treatment process, WMPs, oil, and dispersants will interact by forming WMPs-oil-dispersant agglomerates (WMODAs) (Yang et al., 2021).

When heavy oil and WMPs meet, they will attract each other because of hydrophobic tails and are further wrapped by dispersants. Hence, WMODAs are formed on the crust of dispersants and the core of WMPs and oil. The formation of WMODAs can affect the transportation of WMPs since the oil from WMODAs is less dense than water. Moreover, WMODAs can also impact the oil droplet size and efficiency of dispersant on spilled oil. Such impacts are also affected by the weathering degree of WMODAs. Therefore, distinguishing the weathering degree of WMODAs is essential to understand better their impact on oil spill response options such as dispersant application (Yang et al., 2021).

However, in the research topics of weathering dispersed oil and WMODAs, the application of ML is rare to see, and its potential to aid research in these topics needs to be cultivated. Different ML strategies are applied based on the different data characteristics from the two topics, integer values for weather dispersed oil and images for WMODAs. In the former topic, the data are collected from gas chromatography and mass spectrometry (GC/MS) tests on the sample oil. The diagnostic ratios between biomarkers are calculated from 0 days to 60 days in weathered chemically dispersed oil (CDO) and weathered crude oil (WCO) (Song et al., 2019). Several ML algorithms are implemented to detect the patterns of WCO and CDO, including k-nearest neighbor

(KNN), support vector classifier (SVC), random forest classifier (RFC), decision tree classifier (DTC), logistic regression classifier (LRC), and ensemble vote classifier (EVC). The best-performed ML model is used to predict spilled oil type. In the latter topic, the images of WMODAs are acquired from scanning electron microscopy (SEM) (Yang et al., 2021). Deep learning (DL) is applied to detect differences between different weathering degrees of WMODAs images. Different DL architectures are performed to achieve the best accuracy in the classification, including customized convolutional neural networks, residual networks (Resnet), visual geometry group (VGG), and transfer learning. The developed DL model can classify different weathering degrees of WMODAs with high accuracy. Therefore, the application of ML is proved efficiently to solve the challenges stated above and has a general reference to another environmental research.

1.2 Statement of problems

As introduced in the background section, ML and DL still face challenges in training in some domains. The ML algorithms tend to overfit sequential datasets with hundreds and thousands of features if they are trained directly without preprocessing. The overfitting can be incurred by a high correlation of some features that can cause the training process's abundance. This phenomenon is more evident in the oil fingerprinting field since it could provide hundreds or thousands of diagnostic ratios. Without data preprocessing, the classification would overfit, and it also adds difficulty in the deployment since the data input process is cumbersome. In the image domain, convolution is one of the fundamental methods in DL, and the deep neural networks based on convolutional layers lack explanation. The low expandability leads to low reliability and further reduces models' application. Hence, appropriate explanation algorithms could be considered to

integrate with deep neural networks to increase models' reliability. In some domains, the features detected by deep neural networks are subtle to humans, and the explanation algorithms can highlight the region of the predicted image class to increase the understandability of how the model predicts. Apart from classification, DL also shows strong ability in data augmentation. DCGANs have been regarded as a promising data augmentation method to increase the size and generalization of datasets. However, it faces some challenges, such as mode collapse due to binary cross-entropy loss function and vanishing gradient due to the overconfidence of discriminators (Bang and Shim, 2021). In order to acquire high-quality generated images, the optimization for a stable training process is much needed.

In marine oil spill response, few existing modeling methods have been reported for source identification in oil fingerprinting to distinguish WCO and CDO. It is essential to differentiate WCO and CDO because the change is happening in diagnostic ratios between WCO and CDO, and it could cause an error in the source identification through the database. Traditional statistic methods are not ideal in handling hundreds or thousands of features because of high interactions between each feature. Moreover, they cannot perform classification programmatically, particularly in continuously improving accuracy by updating data.

The recent advancement of ML methods presents great potential in improving oil fingerprinting. The developed ML models can better classify different types of oil and conduct a matching process with the source oil database. Hence, it can fill the gaps and support more accurate and reliable classification and identification of spilled oils in environmental forensic and oil spill response.

Furthermore, the existence of MPs in the marine environment and the interactions with oil (and dispersed oil) forming WMODAs can affect both the transport and fate of MPs and the

degradation and dispersion effectiveness of spilled oil. The effects that WMODAs incur usually correlate to the weathering degrees. Distinguishing the weathering degree of WMODAs is essential to evaluate their impact on response options such as dispersant application and support decision-making in marine oil spill response operations. However, there is lacking data/image analysis efforts in classifying WMODAs under different weathering conditions, particularly with limited observed datasets. Data augmentation becomes valuable to handle data scarcity, prevent overfitting, and improve accuracy.

1.3 Research objectives

To help address the above challenges, the main objectives and tasks of this dissertation research are as follows:

(1) To develop an approach based on meta-ML algorithms and integrate with principal component analysis (PCA) to preprocess datasets by eliminating high correlated features and reconstructing new features. This dimensionality reduction algorithm can reduce bias and facilitate the data input for the prediction, and the model performance is further evaluated by confusion matrix and f score. The developed approach is further tested on the first case study of the classification of sequential data of WCO and CDO in oil fingerprinting. Diagnostic ratios from five different biomarkers are selected as data input. Different ML algorithms are used to predict class labels, and comparative analysis is conducted to choose the most accurate model for the classification task, and it is then deployed for general public use.

(2) To develop an integrated approach of convolutional neural networks (CNNs) and optimized DCGANs. The optimized DCGANs are enhanced by mainly three techniques

for better stability, e.g., label smoothing, spectral normalization, and noisy labels. Through the data augmentation of optimized DCGANs, DL algorithms can be freed from the limitation of datasets. To further strengthen the model's reliability, local interpretable model-agnostic explanations (LIME) are supplemented to interpret prediction from the CNNs model. The CNNs model can also be improved from the quality of interpreted results. The developed approach is further applied in the task of the classification of image data of WMODAs under different weathering degrees. SEM images are taken as data input. CNNs are tailored for the shape and channels of the image of SEM images. The loss and the quality of generated images are the evaluation standards for the performance of optimized DCGANs, and the model accuracy will be evaluated for CNNs performance. Apart from self-designed CNNs, transfer learning models based on CNNs are also introduced to the classification and further compared with selected models for better performance.

1.4 Structure of the thesis

This thesis consists of five chapters. Chapter 1 outlines the general research background and scopes, research objectives, and thesis structure. Chapter 2 provides the literature reviews of the relevant topics, including (1) current widely used ML/DL algorithms in supervised learning and unsupervised learning and the challenges and limitations they face, (2) related research and challenges of oil fingerprinting in oil spills and MPs in marine pollution, (3) potential application of ML/DL on the challenges in these two areas. Chapter 3 presents the development of the binary-classification approach using diagnostic ratios from different biomarkers and illustrates its application for a case study on oil fingerprinting analysis. To further support in-depth analysis of environmental data, images containing complex data are usually used, leading to more powerful

convolutional neural networks algorithms. Chapter 4 describes the interpretation-orientated deep learning method for data classification based on environmental images. The developed approach is implemented to classify weathering degrees of WMODAs using SEM and investigate the impact of MPs on dispersed oil to aid oil spill responses. Finally, Chapter 5 concludes this research with recommendations for future work. The structure of the thesis is illustrated in Figure 1.3.

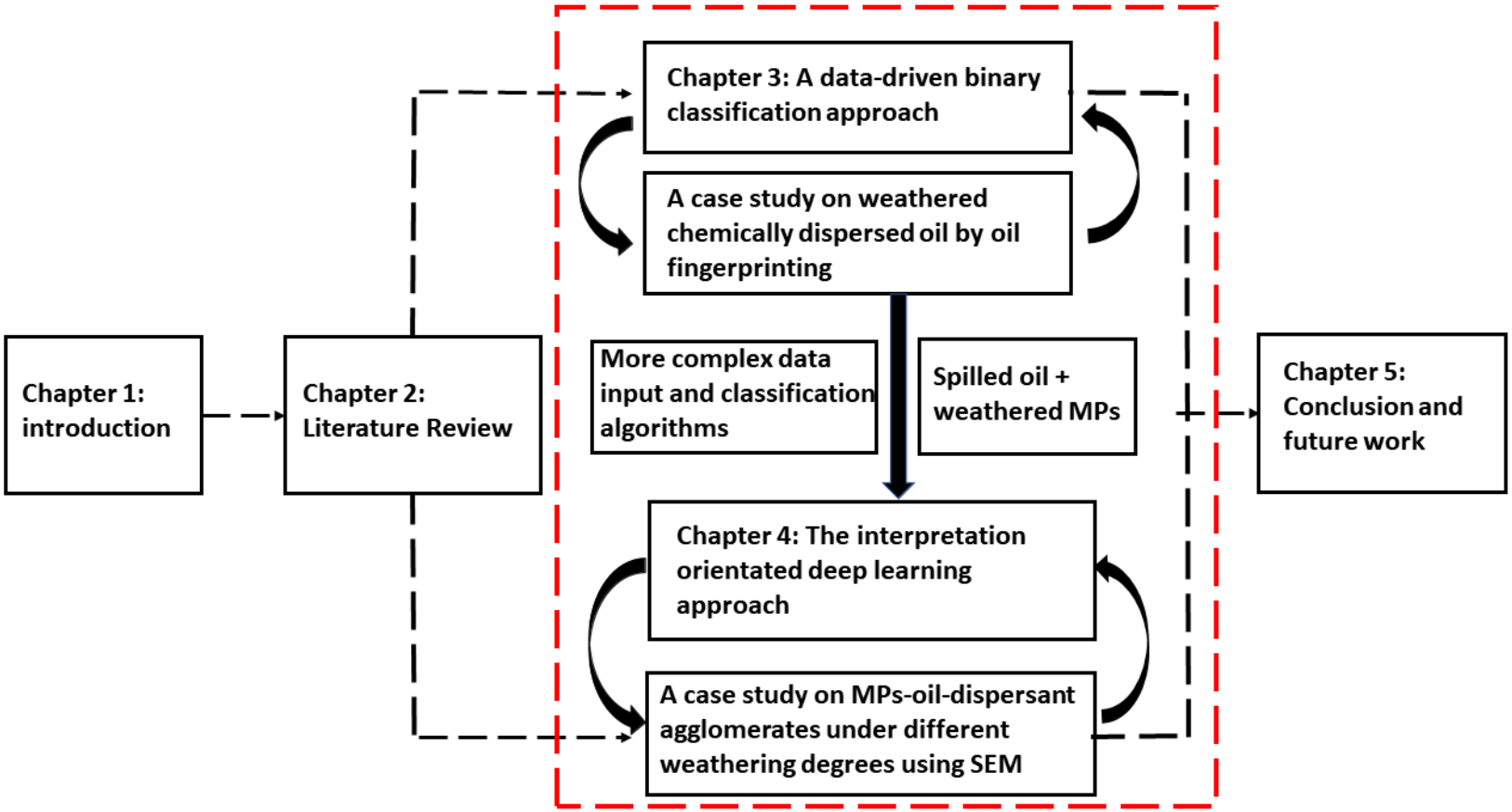


Figure 1.3 Roadmap of the research

CHAPTER 2: LITERATURE REVIEW

2.1 Machine learning

2.1.1 Machine learning techniques

As a central component of AI, ML has been reorganized as a separate field and started to flourish in the 1990s (Kristian 2018). There are many definitions of ML, but in short, ML is a form of AI that allows software programs to improve their prediction accuracy by detecting patterns in the input without being expressly designed in that way (Goodfellow et al., 2016). In ML, there are two kinds of data: labeled and unlabeled data. Labeled data includes both the input and output parameters in a machine-readable manner. However, labeling the data needs a significant amount of human effort. Unlabeled data contains only one or no parameters in machine-readable form (Bach et al., 2017). This eliminates the need for human labor but necessitates more complicated solutions. There are three major machine learning algorithms: supervised learning, unsupervised learning, and reinforcement learning (Ayodele, 2010). The former two types will be introduced mainly and applied in environmental data.

Supervised learning is one of the most fundamental forms of ML. The ML algorithm is trained on labeled data in this type. Even though the data must be appropriately labeled for this approach to operate, supervised learning is incredibly effective when utilized in the right conditions. The supervised learning algorithms are given a short training dataset as a subset of the larger dataset. This short training dataset helps to provide the algorithm with a rudimentary understanding of the issue, solution and data points to be handled (Nasteski, 2017). The training dataset is also quite close to the final dataset in terms of properties, and it supplies the algorithm with the labeled parameters necessary for the task. The algorithm then discovers correlations between the parameters provided, effectively constructing a cause-and-effect link between the

variables in the dataset. After training, the algorithm understands how the data works and the relationship between the input and the output. This solution is subsequently deployed for usage with the final dataset, from which it learns in the same manner as it did with the training dataset (Jiang et al., 2020).

Some of the most commonly used algorithms for supervised learning are linear regression, support vector machine, and logistic regression. A linear regression model attempts to fit a regression line to the data points that accurately capture the relationships or connections. Ordinary least squares (OLE) is the most commonly used approach. The optimum regression line is obtained using this approach by minimizing the sum of squares of the distance between data points and the regression line (Kong et al., 2020). The example of how the regression line is drawn in a dataset is shown below in Figure. 2.3.

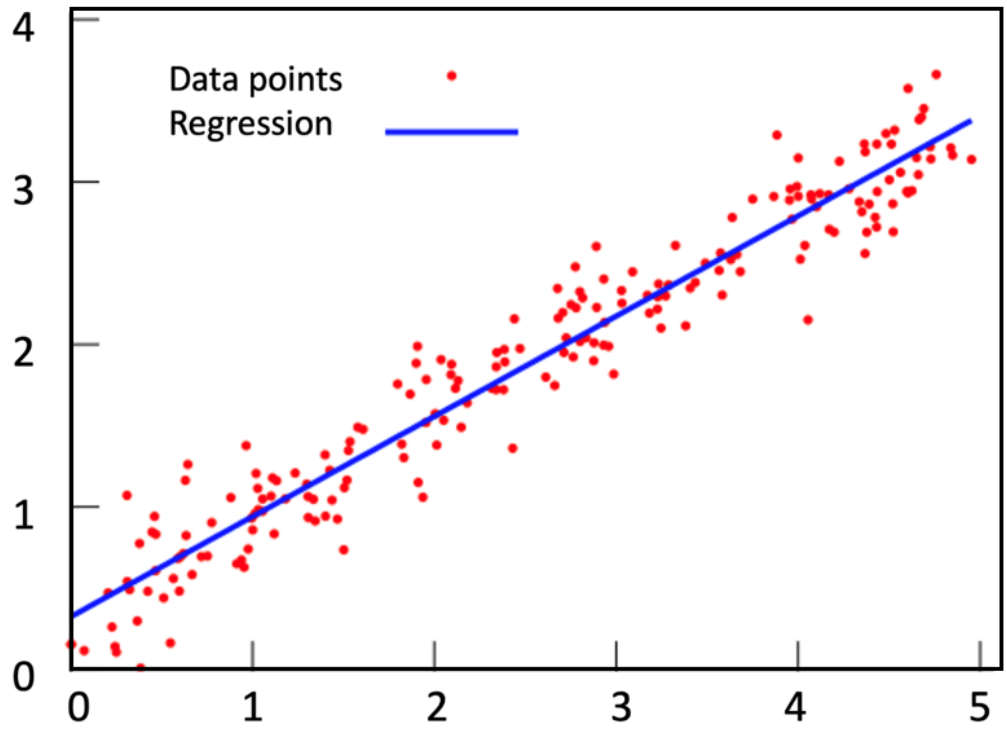


Figure 2.1 Demonstrate of linear regression's decision boundary

SVM differentiates between classes by constructing a decision boundary. Each observation (or data point) is displayed in n-dimensional space before establishing the decision boundary. The number of features utilized is denoted by "n." The decision boundary is designed to maximize the distance to the support vectors. If the decision boundary is too close to a support vector, it will be susceptible to noise and fail to generalize appropriately. Even minor changes in independent variables might result in misclassification (Golbayani et al., 2020). SVM is extremely useful when the number of dimensions exceeds the number of samples. SVM finds the decision boundary by using a subset of training points rather than all, saving memory. On the other hand, training time increases for big datasets, which negatively impacts performance (Sheykhmousa et al., 2020). Figure 2.4 shows how the decision boundary is drawn by maximizing the margin between support vectors.

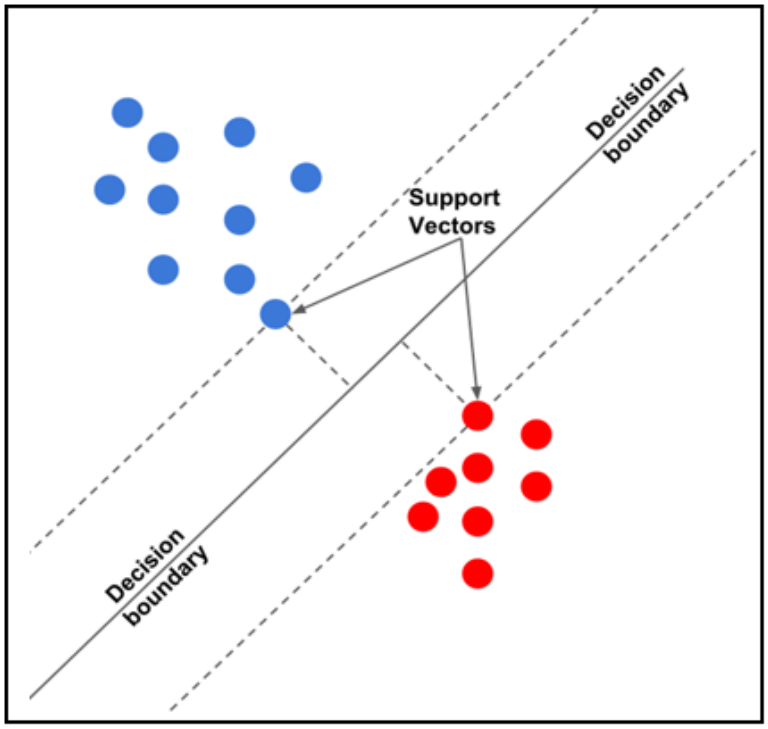


Figure 2.2 Decision boundary from SVM (Cura, 2020)

The logistic regression model takes a linear equation as input and performs a binary classification task using a logistic function and log odds. The logistic function, commonly known as the sigmoid function shown below in equation 2.1, is the foundation of logistic regression. It takes any real-valued integer and translates it to a value between 0 and 1 (Guo et al., 2020).

$$y = 1 / 1+ e^{-x} \quad (2.1)$$

where x is data input; y is the output of the algorithm.

Unlike supervised machine learning, unsupervised machine learning benefits working with unlabeled data. This implies that no human labor is necessary to make the dataset machine-readable, allowing the software to work on a much bigger dataset. Labels in supervised learning allow the algorithms to determine the exact nature of the link between any two data points (Kim et al., 2020). On the other hand, unsupervised learning lacks labels, leading to the information of hidden structures. The program perceives relationships between data points abstractly, with no human input necessary. The construction of these hidden structures gives unsupervised learning algorithms their versatility. Instead of a predefined and fixed problem statement, unsupervised learning algorithms may adapt to the input by constantly modifying hidden structures. This provides greater post-deployment development than supervised learning techniques (Zhong and Leonard, 2020).

Unsupervised learning algorithms include clustering, anomaly detection, neural network, and others. Clustering is a crucial concept in unsupervised learning. It is primarily concerned with identifying a structure or pattern in a set of uncategorized data by finding natural clusters (groups). Types of clustering include K-means clustering a, principal component analysis (PCA), and others (Grira et al., 2004). K-means is a centroid-based or distance-based technique that computes

distances to assign a location to a cluster. Each cluster in K-means is paired with a centroid. The K-means algorithm's primary goal is to minimize the sum of distances between points and their corresponding cluster centroid (Sinaga and Yang, 2020). PCA is an unsupervised statistical approach for dimensionality reduction. It decreases the number of associated variables into fewer independent variables while retaining the essence of these variables. It provides an overview of the linear connections between inputs and variables. PCA can also improve high-dimensional data visualization by constructing new features from the original dataset. The dimensionality reduction feature realizes the data visualization in two or three dimensions and brings more insights into datasets before further training (Kumar et al., 2017).

2.1.2 Data preprocessing

Data preprocessing is the process of converting raw data into a comprehensible format. It is also a critical stage in data mining since raw data cannot be directly worked with. Before deploying machine learning or data mining techniques, the data quality should be evaluated (Huang et al., 2015). The data quality evaluation can be followed as shown in Table 2.3 (Kotsiantis et al., 2006).

Table 2.3 Criteria to evaluate data quality

Criteria	**Content**
Accuracy	Determining whether the data entered is correct.
Completeness	Checking whether the data is available or not recorded.
Timeliness	The data should be updated regularly.

Interpretability	The understandability of the data.
Consistency	Check if the same data is retained in all locations that match or do not match.

In order to realize these criteria, there are four primary tasks to be conducted: data cleaning, data integration, data reduction, and data transformation, as shown in Figure 2.5 (García et al., 2015).

Figure 2.3 Four significant aspects of data preprocessing

Data cleaning is the process of removing erroneous, incomplete, and inaccurate data from datasets and replacing missing information. When handling noisy data, there are three primary methods: binning, regression, and clustering (Chu et al., 2016). Binning is used to smooth or deal with noisy data. The data is first sorted and segregated before storing in bins. Smoothing data in the bin can be accomplished using one of the three approaches: Smoothing with the bin mean, median, and boundary methods (Krishnan et al., 2016). In the first and second ones, the values in the bin are replaced by the bin's mean and median values. In the last approach, the lowest and maximum values of the bin values are obtained and replaced with the nearest boundary value. Regression aids data handling when there is extraneous data and helps determine which variables are appropriate for analysis. Clustering is used to discover outliers and group data (Ilyas and Chu, 2019).

Data integration is the process of merging data from several sources into a single dataset, and it is one of the essential aspects of data management. There are three problems to be considered during data integration: schema integration, entity identification problem, and detecting and resolving data value concepts (Li et al., 2018). Schema combines metadata from many sources. The challenge of entity identification is to identify entities from numerous databases. The detection and resolution of data value concepts need to consider the data taken from different databases while merging may differ (Dong and Rekatsinas, 2018).

Data reduction aids in the decrease of data volume and storage space, making analysis easier while producing the same or almost the same results. Dimensionality reduction, numerosity reduction, and data compression are examples of data reduction approaches (Czarnowski and Jędrzejowicz, 2008). The purpose of dimensionality reduction was illustrated in PCA above, and it helps reduce storage space and computation time. Numerosity reduction reduces the amount of

the data to make it smaller by choosing suitable forms of data presentation while no data loss occurs. Data compression refers to the process of compressing data by encoding, reconstructing, or modifying data (Ougiaroglou et al., 2018).

Data transformation refers to changing the format or structure of data. Depending on the requirements, the difficulty of data transformation varies (Jiang et al., 2008). Typical techniques for data transformation are smoothing, aggregation, discretization, and normalization. Data smoothing can eliminate noise from the dataset by using algorithms and discovering minor changes that aid in prediction. Data aggregation keeps and displays data in summary with a data analysis description. It is a critical phase to ensure the data quality and quantity. Discretization divides the continuous data into intervals and reduces data size (Kara et al., 2018).

2.1.3 Strategies of data augmentation

Data augmentation is a method of creating additional training data from existing training data. It is accomplished by applying domain-specific approaches to examples from the training data, resulting in new and distinct training instances, which helps increase the size of the dataset and introduces variability in the dataset. By creating a larger dataset, models will be better at generalizing to circumstances they may encounter in production (Mikołajczyk and Grochowski, 2018). For example, introducing random noise in a self-driving car dataset may make the model more resistant to camera errors (Jöckel et al., 2019). Similar scenarios also occur in the environmental domain. For example, distinguishing microplastic images from other micro particles will also be beneficial from introducing random noise to simulate complicated situations in reality. In the training of deep neural networks, large datasets cannot handle underfitting but only overfitting. Data augmentation can be used on any data, e.g., image, audio, text, and others

(Aggarwal, 2019). Image data augmentation is mainly introduced since images exist in the environmental domain. There are mainly geometric transformation methods in the transformation of image data, such as rotations, shearing, changes in scale, translations, horizontal and vertical flips, and others. After applying geometric transformation augmentation, the model can be trained only by augmented data or both original and augmented data (Zoph et al., 2020).

Except for geometric augmentation methods, cutting-edge augmentation methods use generative adversarial networks (GANs). The GANs model architecture includes two sub-models: a generator model for producing new instances and a discriminator model for identifying whether created examples are real, from the domain, or fake, generated by the generator model (Luo and Lu, 2018). The generator model creates a sample in the domain using a fixed-length random vector as input. A Gaussian distribution is employed to generate the vector and then seed the generative process. Points in this multidimensional vector space will correspond to points in the issue domain during training, resulting in a compressed representation of the data distribution. A latent space, or a vector space containing latent variables, is the name given to this vector space (Lim et al., 2018; Waheed et al., 2020). The discriminator takes a domain example (real or generated) as input and predicts whether it is authentic or fake. The real example is taken from the training dataset, and the generator model produces the generated examples. During training, the loss function in the discriminator will calculate the loss between generated and actual examples and adjust the parameters in the network by back-propagation (Lim et al., 2018). After training, the discriminator is discarded since the trained generator is mainly used to generate examples. The fundamental GANs structure is illustrated in Figure 2.6 (Vallecorsa et al., 2019).

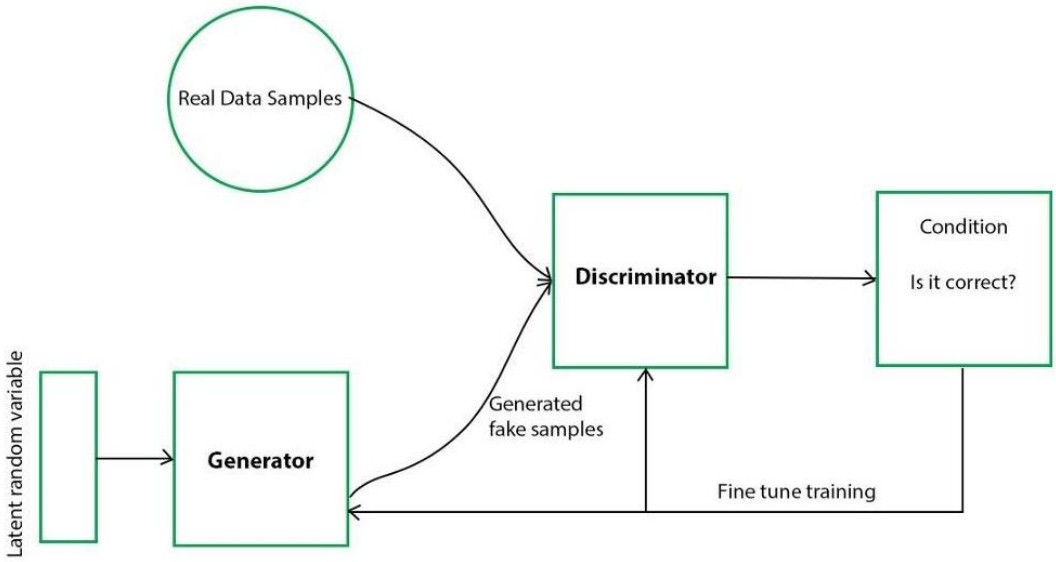


Figure 2.4 Fundamental architecture of GANs (Vallecorsa et al., 2019)

2.1.4 Applications in the environmental field

Because of its tremendous fitting abilities, ML has shown promise in tackling complicated data patterns or formats. As a result, ML, particularly deep learning, has seen significant growth in a range of applications over the last decade, including image classification and machine translation. Researchers in the broad area of environmental science and engineering (ESE) have eagerly embraced ML as well in many applications, for example, assessing environmental hazards (Tollefson et al., 2021), evaluating the health of water and wastewater infrastructure (Granata et al., 2017), improving treatment methods (Inoue et al., 2017), detecting and characterizing pollutant sources (Huang et al., 2021), and performing life cycle analysis. Compared with traditional statistical tools (Gao and Pishdad-Bozorgi, 2020). ML is especially suitable for solving complex environmental problems for the following reasons. ML has the capacity to address a large number

of factors that have weak or nonlinear correlations with the results. Furthermore, in situations where the critical information is not contained in a single input variable, nor are the essential variables known ahead of time, ML can be more effective than traditional statistical tools in handling various data formats, such as text, images, and graphs, where some previously unknown combination of features is required to determine the outcome (Bini, 2018). The general process of ML application in the environmental domain is shown below in Figure 2.7.

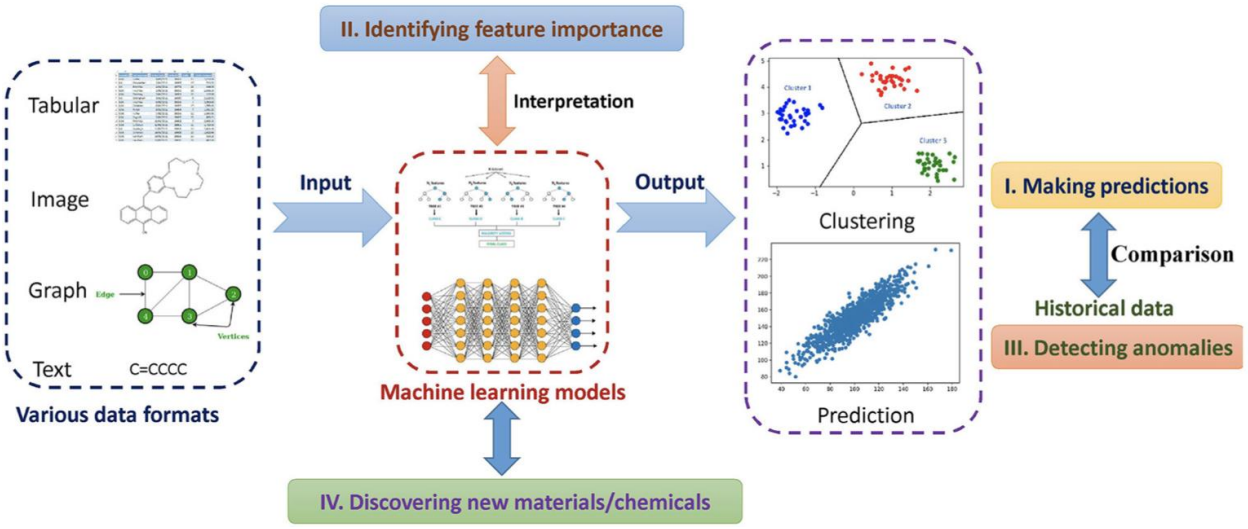


Figure 2.5 General process of ML application in the environmental domain (Zhong et al., 2021)

Three examples are introduced below to illustrate better how ML can address different environmental problems.

The first example focuses on the MPs. Due to MPs' potential influence on water pollution, wildlife, and the food chain, MPs have gained significant attention these years. After sieving and digestion, reliable, quick, and high-throughput screening of MPs from other components of a water sample remains a highly desirable aim to avoid time-consuming visual inspection under the optical microscope. Vittortio et al. (2019) proposed a novel technique that combines 3D coherent imaging

with ML to accomplish accurate and automated identification of MPs in filtered water samples at the microscale. During the water pre-treatment procedure, sediments and aggregates that fall outside the measured range are removed. However, it is still crucial to differentiate MPs from marine microalgae. It is demonstrated that by creating a distinct collection of different "holographic properties," MPs within the prescribed analytical range may be reliably identified.

The second example is from the air pollution domain. Suleiman et al. (2019) introduced a new way for assessing the efficacy of roadside PM10 and PM2.5 reduction scenarios using ML-based models, including artificial neural networks (ANN), boosted regression trees (BRT), and SVM. The ML models predicted PM10 and PM2.5 concentrations well, with around 95% of predictions coming within a factor of two of the actual values at the roadside.

The last example is a water quality-related problem. Lu et al. (2020) proposed two innovative hybrid decision tree-based ML models to generate more accurate short-term water quality predictions. Extreme gradient boosting (XGBoost) and random forest (RF) were the base models of the two hybrid models, which both added an advanced data denoising approach. Six water quality indicators, including water temperature, dissolved oxygen, pH value, specific conductance, turbidity, and fluorescent dissolved organic matter, were predicted using two hybrid models. Their results revealed the low mean absolute percentage errors in predicting temperature, dissolved oxygen, and specific conductance.

These three examples present the most applications of ML in the environmental domain as classification and regression problems with the different input data types, ranging from image and tabular inputs. It shows the tremendous flexibility of ML applications in dealing with different environmental problems. However, in the oil spill and marine pollution fields, ML models are not

as widely applied as they are in the mentioned fields. The following two sections will illustrate current problems and challenges in oil fingerprinting in oil spills and MPs in marine pollution, and in the following two chapters, the proposed approaches are demonstrated how ML can be used to solve these problems.

2.2 Oil Fingerprinting

2.2.1 Marine oil spills and oil fingerprinting

Petroleum products are increasingly produced and consumed as the world's primary energy source, and their environmental effect is also growing. The possibility of significant oil spills persists, despite tremendous progress in decreasing leakage through a combination of technological and regulatory preventative methods and improved business practices. Hundreds to thousands of spills are estimated to occur every day worldwide, including various types of crude oil to a wide range of refined products, from heavy, long-lasting fuels to light, short-lasting, but very poisonous fuels in the marine environment. The fate, behavior, and influence of spilled oil in the marine environment are determined by the spilled oil's chemical composition and bulk properties and the accompanying weathering process. Marine oil spills are a significant source of worry due to the enormous financial costs and persistent, severe damage to the marine ecosystem, local economy, and coastal society. There are ten effective methods for cleaning up marine oil spills: using oil booms, skimmers, sorbents, burning in-situ, dispersants, hot water, high-pressure washing, and manual labor bioremediation, chemical stabilization of oil by elastomers, and natural recovery. Each method has different advantages and disadvantages and is appropriate to use on different occasions. For example, oil booms only work when the oil is in one spot or calm marine environment, and skimmers are then applied in the confined area to separate the oil from the water.

When in the scenario that booms cannot contain spilled oil, then dispersants are applied to accelerate the disintegration of oil. They increase the surface area of each molecule, allowing the oil to chemically bind with water and prevent the slick from spreading across the water's surface, and making it easier for microbes to break down the oil.

In oil spill response, cleanup methods are not the only concern needed to be considered. Determining the source of spilled oil, differentiating and correlating oils, and monitoring the degradation process and weathering state of oils under various conditions are all critical steps to take. Furthermore, allocating legal liability is also essential for oil spill recovery. For the above responsibilities, oil fingerprinting analysis is an effective solution to address them. Oil fingerprinting is one of the critical technologies to describe procedures that use geochemical analysis of the composition of hydrocarbon fluids spilled into the environment (Wang et al., 2006). During the formation of oil and gas, hydrocarbons are affected by many processes, such as biodegradation, gas flushing, water washing, and evaporation.

Furthermore, temperature, reservoir compartmentalization, aquifer activity, and other variables influence the degree of change. Subsequently, hydrocarbons that originated in one source rock have distinct properties in other reservoirs (Mulabagal et al., 2013). Identifying hydrocarbons from different reservoirs can be realized by analyzing variations in composition or identifying unique "fingerprints" of hydrocarbons. Oil fingerprinting techniques have been frequently used to establish the sources of an oil spill by comparing compositional characteristics of both the spilled oil and probable sources. Identifying oil sources and their characteristics is critical for determining the spill's fate and environmental impact, providing a suitable spill response, and assigning obligations and liabilities (Song et al., 2016).

2.2.2 Traditional methods in oil fingerprinting

Many advanced instruments for identifying biomarkers have recently become available, including comprehensive two-dimensional gas chromatography, and mass spectrometry (GC/MS), isotopic resolution mass spectrometry (IRMS), electrospray ionization liquid chromatography-mass spectrometry (ESI-LC-MS), and ultrahigh-resolution Fourier transform ion cyclotron resonance mass spectrometry (FT-ICRMS) (Mansuy et al., 1997) (Cho et al., 2012). In recent years, GC/MS has been a well-known approach with a better-resolving capacity to separate groups of hydrocarbons, particularly isomers and hydrocarbons with identical retention times. Based on GC/MS, the impacts of physio-chemical weathering on biomarkers can be effectively examined (Yang et al., 2011).

Diagnostic ratios are the primary indications for oil fingerprinting, allowing the tasks of oil source identification, characterization, and weathering tracing to be realized. A significant advantage of the diagnostic ratio of spilled oil and suspected source oil is that the influence of the concentration is reduced, and the use of ratio tends to produce an autonomously normalizing effect on the data (Song et al., 2018). Usually, a quantitative oil source fingerprinting is conducted by GC/MS methodology, and specific diagnostic biomarker ratios are determined using many published combinations of ratios and a few new ratios using similar guidelines. There are eight major biomarker classes, and each of them has different diagnostic ratios. The following table gives examples of diagnostic ratios under each biomarker (Song et al., 2019).

Table 2.1 Common diagnostic ratios from biomarkers(Song et al., 2019)

Biomarker classes	**Diagnostic ratios**
Acyclic isoprenoids	pristine/phytane
	pristane/n-C17
Terpanes	TR23/TR24
	C23 tricyclic terpane/C30α β hopane
	C24 tricyclic terpane/C30α β hopane
	C24 tertracyclic/C26 tricyclic (S)/C26 tricyclic (R) terpane
	C2718α, 21β -trisnorhopane/C27 17α, 21β -trisnorhopane
	C28 bisnorhopane/C30α β hopane
	C29α β -25-norhopane/C30α β hopane
	C29α β -30-norhopane/C30α β hopane
	oleanane/C30 α β hopane
Steranes	C27ββ(S+R)/C29β β (S+R)
	C28 α β β/C29 α β β steranes (at m/z 218)
	C27 α β β/(C27 α β β +C28α β β +C29α β β) (at m/z 218,
	C28 α β β/(C27 α β β +C28 α β β +C29 α β β) (at m/z 218)
	C29 α β β/(C27 α β β + C28 α β β +C29 α β β) (at m/z 218)
	C27,C28, and C29 α α α/α β β epimers (at m/z 217)
Sesquiterpanes	Peak 5/Peak 3
	Peak 10/Peak 3
	Peak 1/Peak 3

Continued Table 2.2 Common diagnostic ratios from biomarkers(Song et al., 2019)

Biomarker classes	Diagnostic ratios
Diamondoids	1,4-DMA, cis/1,4-DMA, trans
	Dimethyl admantane index: 1,3-DMA/(1,3- + 1,4- + 1,2-DMA) 1,3,4-DMA, cis/1,3,4-DMA, trans
	Trimethyl admantane index: 1,3,4-DMA, cis/(1,3,4-DMA, cis + 1,3,4-DMA, trans)
	Ethyl admantane index: 1-EA/(1- + 2-EA)
	Methyl-diamantane index: 4-MD/(1- + 3- + 4-MD)
	Relative distribution of diamantanes: C0-D:C1-D:C2-D:C3-D
Triaromatic steranes	C26 TA (20S)/sum of C26 TA (20S) through C28 TA (20R)
	C27 TA (20R)/C28 TA (20R)
	C28 TA (20R)/C28 TA (20S)
	C26 TA (20S)/[C26 TA (20S) +C28 TA (20S)]
Monoaromatic steranes	C27−C28−C29 monoaromatic steranes (MA) distribution.

After deciding several diagnostic ratios, the calculation is proceeded by dividing the peak heights of compounds with the same mass to charge ratio by the number of compounds.

During the calculation, the final series of diagnostic biomarker ratios will be determined using a 5% fixed relative standard deviation (RSD) which means that any diagnostic ratios exceeding the limit are excluded (Wang 2008). Once the sample oil's diagnostic ratios are developed, the same ratios will be calculated for different suspected source oils and statistically compared with sample oil's ratios. The calculated biomarkers and sediments will be classified into four oil source-fingerprinting categories: match, potential match, inconclusive, and non-match

(Song, 2019). The comparative findings will reveal how effective the diagnostic ratio technique can distinguish non-weathered crude oil from a comparable geographical production area.

This approach works because quantitative diagnostic biomarker ratios are typically more resistant to environmental weathering and will show little or no change over time.

2.2.3 Challenges of oil fingerprinting in the weathered dispersed oil

The past experiments showed that diagnostic biomarker ratios of spilled oil are stable to environmental weathering without human intervention. During oil spills, dispersants, which comprise surfactants and solvents, are often used to remediate marine oil spills. It can lower the interfacial tension between oil and saltwater by increasing the formation of tiny, stable oil-surfactant micelles (i.e., oil-in-water emulsion) (Major et al., 2012). CDO, split oil in a water emulsion bridged by surfactants, can stay in saltwater for extended periods if dispersants are used (Song, 2019). In the experiment of He et al. (2016), they showed that during a medium to long term weathering process, the most selected diagnostic ratios of n-alkane, terpanes, steranes, and polycyclic aromatic hydrocarbon (PAHs) from all oil samples changed, and only four ratios remained good stability because their RSD is lower than 5% (He et al., 2016). The potential for oil biomarker weathering will profoundly affect the calculation and subsequent critical difference analysis of diagnostic ratios used for oil source fingerprinting. As a result, a more robust quantitative oil source-fingerprinting approach is necessary since it is vital to assess its environmental effect, select further reaction countermeasures, and gain a better knowledge of CDO destiny and behavior in maritime ecosystems.

2.3 Microplastics and coexistence with oil spills in marine environments

2.3.1 Microplastics in oceans

MPs are defined as particles with a diameter of less than 5 mm created by the breakdown of more considerable plastic trash or by the direct production of tiny-sized particles (He et al., 2016). They can be divided into two main categories by their source: primary MPs and secondary MPs. Primary MPs are the main source directly released in the environment by laundering synthetic clothes, abrasion of tyres when driving, and others. Secondary MPs originate from the decomposition of more significant plastic objects, such as plastic bags, bottles, etc. From the report of primary MPs in the oceans, there are seven major sources of primary MPs: tires, synthetic textiles, marine coatings, road markings, personal care products, plastic pellets, and city dust (Boucher and Friot, 2017). In terms of MPs' pathway to ocean, there are many different routes, such as riverine input, wastewater effluent, sewage disposal, litter coastal activities, litter from marine activities, and atmospheric deposition. After MPs enter the ocean, they can cause an enormous impact on the marine environment. MPs may disrupt the development of marine animals and constitute harm to marine ecology due to their tiny size and widespread dispersion.

Furthermore, MPs have been implicated in transporting organic contaminants such as antibiotics and insecticides, resulting in complicated interactions (Everaert et al., 2020). A recent sorption kinetics investigation on the behavior of crude oil on MPs in both seawater and a simulated fish digestive system indicated the possible function of MPs as a vector for bioaccumulation of hydrophobic organic contaminants (Güven et al., 2017). Therefore, MPs are persistent, and a possible vector of toxic organic compounds into the marine environment, and their negative impact is not only physical but also chemical due to their ability to adsorb and accumulate a variety of contaminants (Michele et al., 2021).

2.3.2 Behaviour and effects of microplastics in oil spill

When MPs are released into the ocean, they can integrate with spilled oil and dispersants due to their physical and chemical properties. The integration can potentially affect many processes, for example, oil dispersion, fate, and transport of MPs and spilled oil. Therefore, studying the interaction between MPs and oil spills is necessary to provide more evidence and insight into the synergy between these two ocean pollutants.

There have numerous oil spill response strategies, as stated in the introduction. Among these strategies, great attention is currently being paid to the use of oil spill treatment agents, such as chemical dispersants, that can efficiently break oil into tiny droplets under various environmental circumstances (Clayton et al., 2020). For example, approximately 7.9 million liters of dispersants were used to treat 780 million crude oils during the Deepwater Horizon spills in 2010 (Bælum et al., 2012). These oil spill treatment agents and spilled oil in the ocean might adsorb with hydrophobic particles with a rough surface structure, such as marine snow. Marine snow is the decaying material from dead animals, plants, fecal matter, sand, soot, and other inorganic dust (Dissanayake et al., 2018). Marine snow provides particles to aggregate with oil and allow oil to be transported from the top of the sea to the bottom, resulting in an increase in oil concentration at the seafloor (Brakstad et al., 2018).

Moreover, the interaction between particles spilled oil and chemical dispersants might impair oil dispersion's efficacy (Zhao et al., 2017). Furthermore, the formed oil-particle-aggregates could significantly impact the destiny and transportation of oil. For example, they enable oil transport from the saltwater surface to the water column, potentially increasing oil concentration in the seawater column and improving oil droplet stability (Irisson et al., 2017; Brakstad et al., 2018).

These aggregates are biodegradable in general, and the particulates in the aggregates are harmless (Rahsepar et al., 2017). Unlike marine snow, MPs have physicochemical and structural features that distinguish them from natural biodegradable particles, such as higher hydrophobicity, reduced degradation capability, and possible toxicity (Porter et al., 2018). As a result, MPs may have distinct effects on the oil dispersion process, influencing the transport and destiny of oil in the marine environment. More crucially, the presence of an amphipathic dispersant in the MPs and crude oil system may alter the efficacy of oil dispersion and the shape, transit, and destiny of MPs in the marine system (Yang et al., 2021).

Previous research has shown that MP aging develops novel hydrophilic functional groups and that changes in their surface shape might affect how they interact with other pollutants. The size of oil droplets and the efficacy of dispersion would be affected by MP aging on a day-to-day basis (Balakrishnan et al., 2019). In Yang et al. (2021) study, they measured oil droplet size and dispersion effectiveness of light oil and heavy oil under the DOR of 1:25 after different aging degrees with 583 mg/L MP samples (i.e., polyethylene). MP aging substantially influences the droplet size of heavy oil but has little impact on light oil. As seen in Figure 2.1 (a), Light oil droplet size remained consistent between 3.35 ± 0.15 and 4.76 ± 1.79 um after 56 days of aging. The droplet size for heavy oil, on the other hand, grew from 9.07 ± 0.50 um on day 0 to 15.49 ± 0.00 um on day 56, with a high value of 20.31 ± 2.43 um on day 21. The presence of hydrophilic functional groups (OH, C=O, COC—) generated with slow MP surface aging may have assisted the adsorption of the dispersant's hydrophilic heads. In contrast, the dispersant's hydrophobic tails easily adsorbed the heavy oil, and this process allowed for the creation and expansion of WMDOAs (Yang et al., 2021).

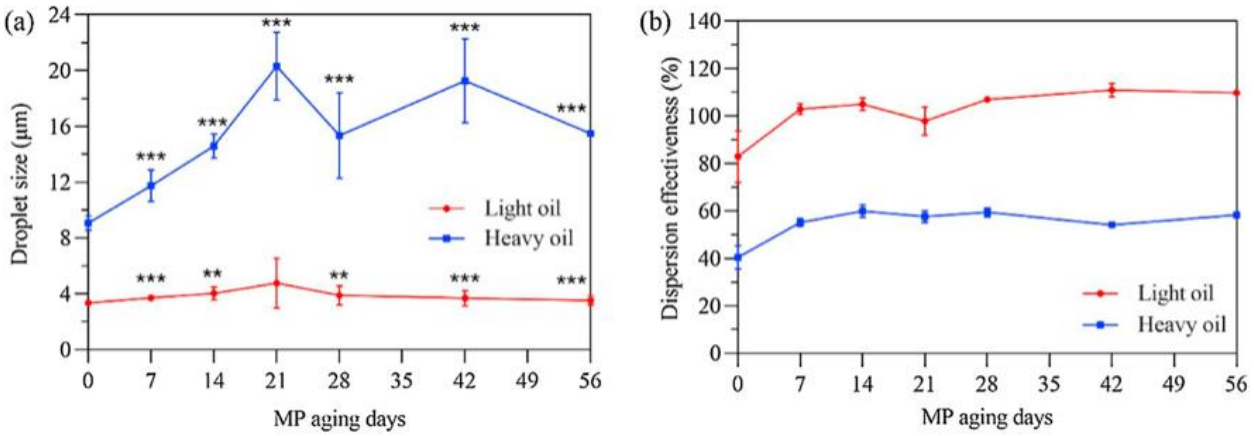


Figure 2.6 the change of droplet size and dispersion effectiveness as MPs aging days increase (Yang et al., 2021)

Figure 2.1 (b) demonstrated that the MP aging process improved light and heavy oil dispersion performance compared to pristine MPs. On day 0, the dispersion efficiency of light and heavy oil was 82.86 ± 10.87 and 40.39 ± 4.96%, respectively, and climbed to 109.75 ± 0.71 and 58.30 ± 0.00%. Aged MPs significantly improved the efficacy of light oil dispersion (about 27%). Aged MPs were more hydrophilic than pristine MPs due to new hydrophilic functional groups (OH, C=O, COC—). As a result, aged MPs might disperse in saltwater with less dispersant usage than pristine MPs. Therefore, more dispersant was available for oil dispersion, increasing the efficacy (Yang et al., 2021). The formation of WMODAs is illustrated below.

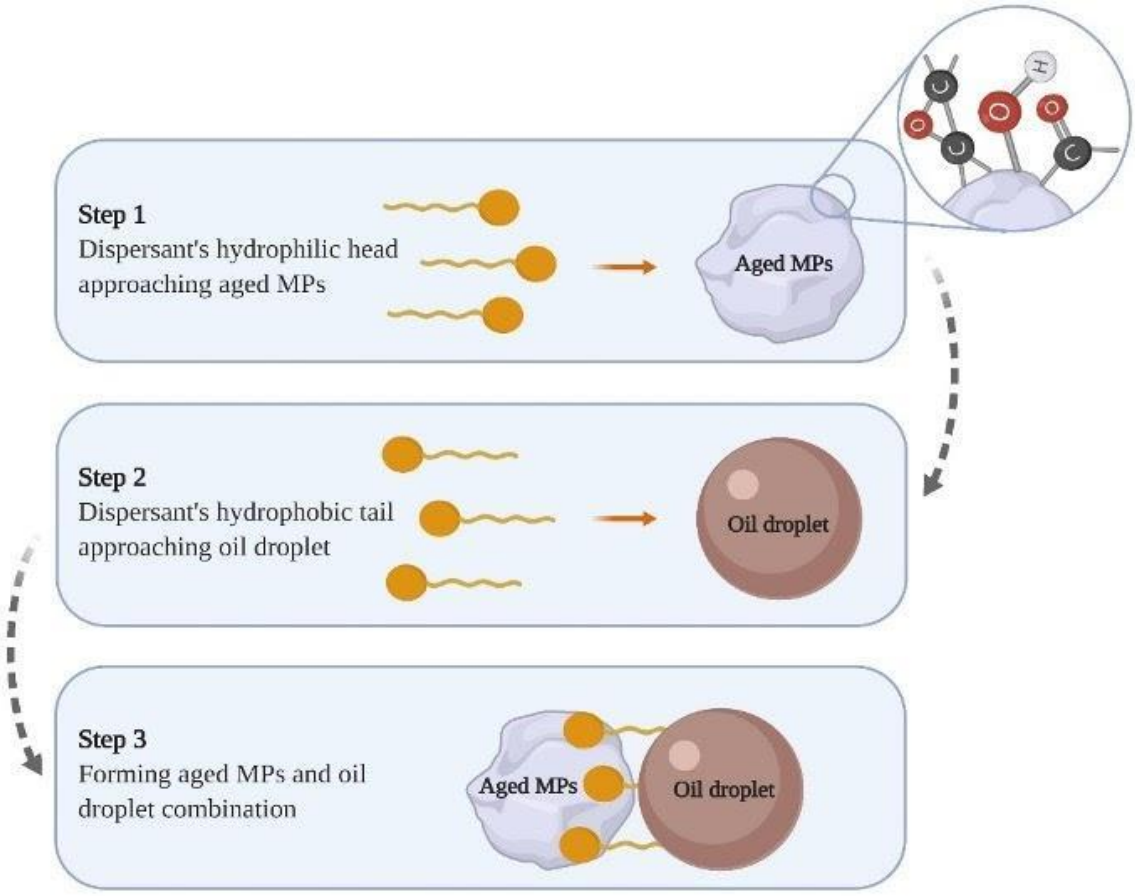


Figure 2.7 Formation process of WMODAs (Yang et al., 2021)

2.3.3 Challenges of distinguishing WMODAs

There are appropriate approaches to identify MPs, ranging from physical to chemical means. Microscopical techniques and chemical analysis (e.g., dissect, polarised, fluorescence, scanning electron, atomic force microscopy, spectroscopy) are the most used methods for identifying micro/nano plastics (Chen et al., 2020; Roch and Brinker, 2017; Shim et al., 2017). The advantages and limitations of the different microscopic and analytic techniques are summarized in Table 2.2.

Table 3.2 Advantages and limitations of analytic techniques (Wirtz et al., 2019; Chen et al., 2020)

Identification method	Advantages	Limitations
Stereo microscopy	Fast and easy. Identification of shape, size, and colors.	Not confirmative of plastic c nature of the particle. Lack of data of transparent or small particles.
Scanning electron microscopy	Clear and high-resolution images of particles. No gas into the chamber if coupled in ESEM mode. Small, detected particles in STEM mode	Expensive. Long time and effort for analysis Lack of information on the type of polymer.
FTIR spectroscopy	Confirmation of the composition of the MPs. No false positive or negative data. Non-destructive analysis of materials	Expensive. Wavelength radiation can be a limiting detection factor. Time consuming to analyze all the particles on a filter

The stereo microscope analyses three dimensions by viewing the sample from two slightly different angles to obtain the two pictures required for stereoscopic vision. Therefore, objects may be observed mainly through reflected light at modest magnifications, generally between 8 and 50 times. In the case of transparent particles, several studies applied stereomicroscope to identify the percentage of plastic-like particles, later characterized by other techniques, roughly 20-70% of the total particles. Furthermore, synthetic and natural fibers (prevalent in water, sediment, and biota samples) are difficult to distinguish using a stereomicroscope (Firdaus et al., 2020). The fluorescence microscope captures fluorescent emission from materials stimulated by a specific wavelength, as opposed to the optical microscope, which depends on the picture's contrast

provided by the reflection of light on the sample. A fluorescence microscope effectively identifies MPs based on their natural propensity to emit fluorescence, mainly white and clear polymers. When paired with imaging, this method lowers MPs detection failure and can lower the size limit of identified MPs (Scircle and Cizdziel, 2019). The quantification of fluorescence spheres with microscopy methods may also identify MPs in various matrices. However, chemical additives in the production process of MPs can potentially alter the fluorescence characteristics and affect the results (Dehghani et al., 2017).

SEM is a method that can provide details on the morphological surface structure of MPs by obtaining high-resolution pictures of the surface state. It can also offer information on the chemical content of the samples because it can be equipped with Energy Dispersive X-ray Spectroscopy (EDS) detectors. The source electrons penetrate the solid material, causing various (both elastic and inelastic) scattering processes to occur, and various detection systems gather the resulting signals to form an image. Several studies used SEM to view MPs in various matrices, including sewage sludge, mussels, sediments, and sand (Anderson et al., 2017; Li et al., 2016; Nguyen et al., 2021).

Infrared spectroscopy (IR) is absorption spectroscopy commonly employed in material characterization to examine chemical bonding. A molecule transition absorbs an infrared photon from its primary to excited vibrational state. Fourier transform infrared (FTIR) spectroscopy is carried out using an interferometer, which allows for the scanning of all frequencies present in the IR radiation produced by the source. FTIR employs four techniques: transmission, reflectance, true specular reflectance/reflection-absorption, and attenuated total reflection (ATR) (Veerasingam et al., 2021). The excellent energy availability leads to a significantly better signal/noise ratio than traditional IR spectroscopy, which is one of the critical advantages that assures improved

performance. FTIR is usually used to characterize MPs. In previous studies, MPs samples are stimulated, resulting in distinct detected vibrations that allow for the generation of a spectrum with a fingerprinting range. The nature of the substance is described by this spectrum, which can be determined by comparing it to established reference spectra. Large particles (>500 nm) can be examined using ATR-FTIR, while tiny particles need micro-FTIR, which allows for simultaneous spectral imaging, mapping, and collecting (Dini et al., 2021) (Gaston et al., 2020). FTIR has been widely employed in MPs research to locate and describe them in sediment, marine species, surface water, and food (Cincinelli et al., 2017; Corami et al., 2020; Harrison et al., 2012).

However, no study has been done in the characterization of WMODAs since it is a relatively new concept in the interdisciplinary area of MPs and oil spills, and the properties of WMODAs also make the characterization difficult, for example, multiple mixed substances, wide-ranged sizes, and subtle change of morphology. From listed identification methods, either one cannot classify WMODAs with different weathering degrees independently. Therefore, there is a need to develop a new approach to identify WMODAs with different weathering degrees and further quantify their impacts on oil dispersion and oil droplet size.

2.4 Summary

In this chapter, section 2.1 reviewed ML's methods and how ML usually conducts classification and regression, and examples that ML was applied to solve complex environmental problems. There have been few ML-related methods in an oil spill and marine pollution, and further, no ML-related methods were developed to classify WCO and CDO and different

weathering degrees of WMODAs. The potential of using ML algorithms with the aid of computer vision and analytic techniques to solve the challenges in these two related problems is promising.

Section 2.2 reviewed the definition, motivations to conduct oil fingerprinting, and current methods of oil fingerprinting. It further illustrated how to select diagnostic ratios from GC/MS through relatively stable RSD and gave examples of commonly used diagnostic ratios. Section 2.2.4 discussed how the weathering process and the application of dispersant on spilled oil could potentially affect oil fingerprinting because of the change in properties of dispersed oil.

Section 2.3 reviewed the problem of emerging contaminants (e.g., MPs) with spilled oil which would bring challenges in oil fingerprinting, oil dispersion, MPs' fate, and transportation in marine environments. The effects of different weathering degrees of WMODAs on oil dispersion efficiency were discussed, and the needs and challenges of distinguishing different weathering degrees of WMODAs were stated.

The following chapters 3 and 4 will give the proposed approaches in an oil spill and marine pollution fields and demonstrate solutions to the problems of distinguishing WCO and CDO in oil fingerprinting and different weathering degrees of WMODAs in spilled oil remediation.

CHAPTER 3: A DATA-DRIVEN BINARY-CLASSIFICATION APPROACH FOR OIL FINGERPRINTING ANALYSIS*

*The chapter is based on a published article: Chen YF, Chen B, Song X, Kang Q, Ye XD, Zhang BY (2021) A data-driven binary-classification framework for oil fingerprinting analysis. *Environmental Research*. 2021(111454). https://doi.org/10.1016/j.envres.2021.111454. Contributions: Yifu Chen - methodology, software, validation, paper drafting and editing; Bing Chen - conceptualization, writing, review and editing; Xing Song - review; Qiao Kang - data curation and review; Xudong Ye - visualization and review; and Baiyu Zhang - review and & editing. All authors have read and agreed to the published version of the manuscript.

3.1 Introduction

In a marine oil spill, source identification and behavior characterization are critical to understand the environmental impact and response of the accident. The subsequent weathering process affects the fate and behavior of spilled oil such as evaporation, emulsification, photooxidation, and biodegradation (Li et al., 2016; Song et al., 2019). Oil fingerprinting is usually achieved by recognizing specific groups of petroleum hydrocarbons, called biomarkers, such as terpanes and steranes (Shen et al., 2020; Wang et al., 2011). These biomarkers have unique distributions in different oil categories, which allows effective fingerprinting based on the diagnostic relationships among biomarkers (Wang et al., 2013).

However, oil fingerprinting becomes more challenging when spilled oil is treated by chemical dispersants (Song et al., 2018). Dispersants can fractionate oil into smaller droplets, theoretically facilitate the biodegradation process, and reduce the exposure to marine animals and accident respondents (Bayable et al., 2021; Lee et al., 2015). Dispersants can significantly affect oil physicochemical properties (e.g., viscosity, boiling point, and iodine value) and oil weathering, and further influences biomarkers' attributes and generate unreliable diagnostic ratios compared with those in non-dispersed oil (Datta et al., 2018; Torres et al., 2020). This consequence could baffle the weathering processes and result in errors after contrasting chemically dispersed oil (CDO) fingerprinting in marine environments (John et al., 2016; Song et al., 2016). Thus, the addition of dispersants could generate non-negligible bias to the tracking of the source and weathering of spilled oil in oceans (Wu et al., 2021).

Traditionally statistical methods, such as cluster analysis, discriminant analysis, and principal component analysis (PCA), have widely implemented in characterizing and tracing spilled and

weathered oil by primarily analyzing spectrum data of biomarkers (Ismail et al., 2016; Mirnaghi et al., 2019). Machine learning (Nasution et al., 2018), as a novel advancement of statistics, has been recently recognized as a promising tool and introduced into environmental fields due to its many advantages (Saha et al., 2016). Firstly, ML is effective in labelled classification problems (Mieth et al., 2016). Secondly, it is specially designed for data with large size and intricate relationship of input variables and samples without depending on validation of the initial assumptions (e.g., whether data follow normal distribution and linearity) (Jordan and Mitchell, 2015). Thirdly, it only requires choosing predictive algorithms by relying on its empirical capabilities without pre-existing knowledge about subjects (Bzdok et al., 2018). There have growing applications of ML in studying complex environmental problems, such as predicting fecal coliform concentrations in wastewater (Khatri et al., 2020), classifying wide-area seafloor habitat by acoustic and visual data (Zelada Leon et al., 2020), and forecasting water quality parameters in coastal waters (Alizadeh et al., 2018). However, few efforts have been reported in employing ML in the area of oil fingerprinting to classify dispersed oil.

This study introduces ML as a new analysis tool by proposing a new binary classification approach to aid source identification in oil fingerprinting by distinguishing weathered crude oil (WCO) and CDO. The approach comprises six ML algorithms and a dimensional reduction algorithm (i.e., PCA), to address the dispersed oil classification problem. The ML algorithms considered in the study include Random Forest (RF), Support Vector Classifier (SVC), K-Nearest Neighbor (KNN), Logistic Regression (LR), Ensemble Voting Classifier (EVC), and Decision Tree (DT). The total 862 diagnostic ratios based on five types of biomarkers (terpanes, steranes, triaromatic steranes or TA-steranes, monoaromatic steranes or MA-steranes, and diamantanes) are chosen from our previous study (Song et al., 2019) as the features for ML and to be further

evaluated by the six algorithms to identify the best algorithm for classification. The approach is expected to provide an efficient analysis method for identifying the WCO/CDO and supporting oil spill source identification.

3.2 Method

3.2.1 Binary classification approach

The binary classification approach comprised seven steps from the front data entry to the application of ML algorithms. The core of the approach was six ML algorithms that have different specializations in analyzing datasets. Figure 3.1 shows the workflow of the approach, where datasets were preprocessed by feature selection, formed different feature sets, trained by ML algorithms, ranked by performance, and finally used to make classifications. Each square represents a machine learning operator, with arrows indicating the direction of the data flow path.

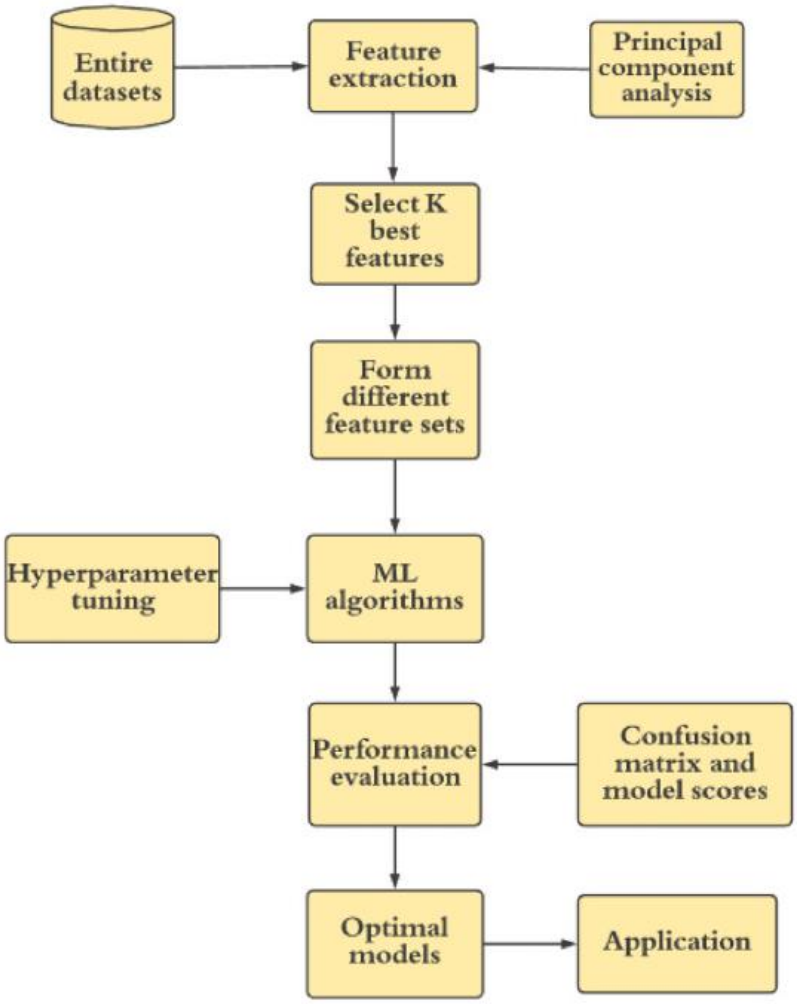


Figure 3.1 The binary classification approach

3.2.2 Data entry and preprocessing

The raw datasets might not be desirable to train ML algorithms directly because of missing data, outliers, or merely heavy computations. In this study, the original datasets were engineered some new features by calculating quotients between every diagnostic ratio. For example, there had up to 462 features in terpanes. This process might also bring noise into the datasets. Without dimensionality reduction technique applied, the input would require 462 values for a single biomarker group. To simplify the feature input process while maintaining the most

information of datasets, principal component analysis (PCA) was introduced in this approach. It decreased hundreds features to dozens, and hence greatly reduced computational time. By reducing dimensionality through PCA, the datasets could be denoised as well.

The main formula of PCA can be represented in Eq. 3.1 (Wetzel, 2017):

$$PC_i = a_1X_1 + a_2X_2 + \cdots + a_dX_d \quad (3.1)$$

where, PC_i, principal component i; X_d, original feature d; a_d, numerical coefficient of X_d.

The Scikit-learn library from Python® v3.8 was used to conduct PCA and data standardization. The 95% variance was selected to keep the information relatively complete in the data to secure accurate prediction results (Hao et al., 2020). Before the feature selection, datasets were normalized to ensure that feature values were simultaneously distributed and contributed equally to the analysis without creating bias (Vasan and Surendiran, 2016). Afterward, the loading matrix in PCA provided the correlations between the original features and new principal components (PC). For better interpretations, feature selection was proceeded by choosing the highest correlated original features (Abdi and Williams, 2010).

3.2.3 Modeling developments for oil fingerprinting

Total six ML algorithms were applied for the comparative analysis. Before feeding input values into ML algorithms, the preprocessed datasets were divided into training sets (80%) and test sets (20%) to evaluate the performance based on previous studies (Bhatnagar et al., 2017; Medar et al., 2017). The brief definitions of six ML algorithms applied in this study were introduced below.

KNN algorithm is a supervised machine learning algorithm. It works by discovering the distances between a query and all the examples in the datasets and choosing the specific number

of examples (K) closest to the query, then voting for the most frequent label. Distance metric could be represented by Euclidean distance in Eq. 3.2 (Potamias et al., 2010):

$$D(x,y) = \sqrt{\sum_{i=1}^{k} \quad (x_i - y_i)^2} \tag{3.2}$$

where, D, the distance between the query and examples; x, the query; y: examples.

SVC algorithm is to find a hyperplane that has the maximum margin between data points of classes in N-dimensional space (N: the number of features) to markedly classified the data points (Wang et al., 2011a). Figure 3.2 demonstrates how the perpendicular distance from the line to the nearest point is used to determine a hyperplane with the greatest margin.

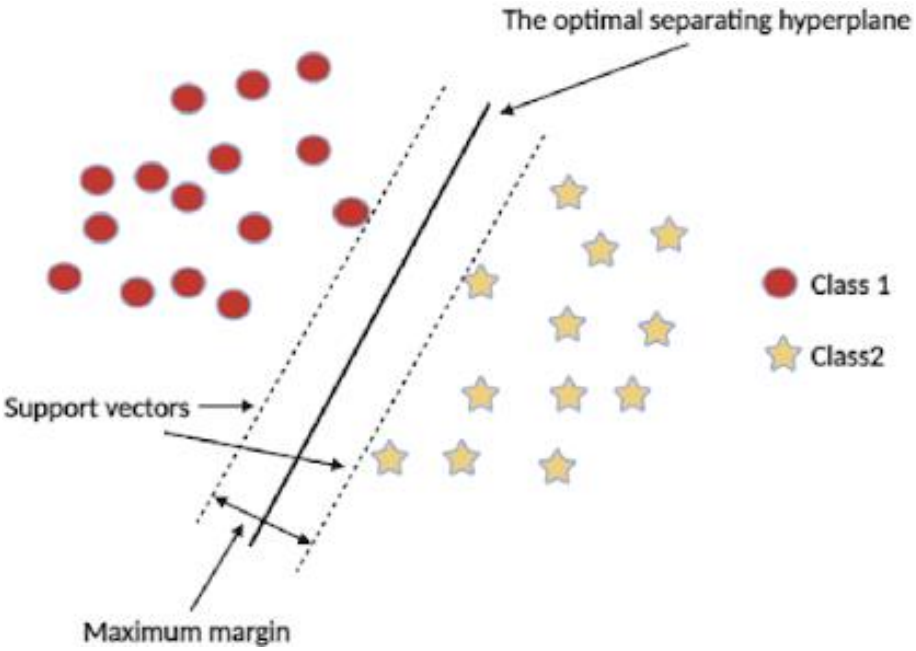


Figure 3.2 The fundamental theory of SVC

DTC consists of nodes, which are the test for the value of a specific edge or branch corresponded to the outcome of a test and connected to the next node or leaf and leaf nodes. Terminal nodes are used to predict the class labels (Farid et al., 2014). RFC is an advanced decision method, which combines the output of multiple (randomly created) decision trees. This ensemble learning increases the classification performance of a single tree classifier by randomly choosing

data nodes to create a decision tree. The feature space is divided into M regions Rm, $1 \leq m \leq M$ by a decision tree with M leaves. The prediction equations (Eq. 3.3 and Eq. 3.4) are defined below for each tree:

$$f(x) = \sum_{m=1}^{M} \quad C_m \Pi(x, R_m) \tag{3.3}$$

where, M, the number of regions in the feature space; R_m, a region suitable to m.

C_m is a constant appropriate to m:

$$\Pi(x, R_m) = \{1, \qquad if\ x\ \in R_m$$

$$0, \qquad otherwise\} \tag{3.4}$$

The final classification prediction come from the majority votes of all trees (Chen et al., 2020; Rodriguez-Galiano et al., 2012).

LR is a statistic model that conducts a logistic function or sigmoid function to model a binary dependent variable, and the logistic curve, which relates the independent variable, can be represented as shown in Eq. 3.5:

$$P_i = \frac{e^{a+bx}}{1+e^{a+bx}} \tag{3.5}$$

where, *P*, the probability of a label 1; *e*, the base of the natural logarithm; *a* and *b* are the model parameters (Robles-Velasco et al., 2020) and the probability threshold used for binary classification is 0.5.

EVC is a meta-classifier that uses majority voting to classify related or potentially different machine learning classifiers, as shown in Eq. 3.6:

$$y = mode\{C_1(x), C_2(x), \ldots, C_m(x)\} \tag{3.6}$$

where, *y*, the predicted label; C_m, different classifiers (Onan et al., 2016).

Depending on the attributes of different ML algorithms, they perform distinctively on the same datasets. Hence it is important to understand each ML algorithm's limitations. Table 3.1 presents their significant advantages and disadvantages.

Table 3.1 Overview of different ML algorithms

Model	Advantages	Disadvantages	References
KNN	Has no training periods. Adds new data seamlessly.	Has high outlier sensitivity. Is unsuited for high-dimensional data.	(Ao et al., 2019; De Leonardis et al., 2018)
SVC	Is suitable for high-dimensional data. Has relatively memory efficient.	Is unsuited for large datasets. Has poor noise resilience.	(Dou et al., 2020; Zendehboudi et al., 2018)
RFC	Deals with very high dimensional data. Has built-in feature importance metric.	Overfits noisy classification problems.	(Ao et al., 2019; Dogru and Subasi, 2018)

DTC	Requires fewer efforts for data preparation.	Has long training process. Is inadequate for regression problems.	(Hamsagayathri and Sampath, 2017; Pham et al., 2017)
LRC	Updates easily to reflect new data. Is less prone to over-fitting in a low dimensional dataset. Extends easily to multi-class classification.	Is unable to solve non-linear problems. Is difficult to capture complex relationships and sensitive to outliers.	(Christodoulou et al., 2019; Lee and Jun, 2018; Qasim and Algamal, 2018)
EVC	Improves overall model performance. Unlikely overfits.	Is hard to interpret ensemble models. Does not perform well when an individual model is closest to the true data generating process.	(Saqlain et al., 2019; Xiao et al., 2018)

3.2.4 Hyperparameter optimization and overfitting

In machine learning, some hyperparameters need to be initialized and adjusted for better prediction results. Meanwhile, overfitting could happen during the model training because models learn the details and noise in the training data to the extent that detrimentally impacts the

performance of models on new data (Yeom et al., 2018). In this approach, GridSearchCV was applied for both hyperparameter optimization and overfitting prevention (Ranjan et al., 2019). The main components of the GridSearchCV are:

- *Hyperparameter grid*

A python dictionary was created with hyperparameter names as keys and a list of hyperparameter settings as values. The best accuracy was verified based on the test on all the combinations of hyperparameters.

- *Cross-validation*

Cross-validation was a statistical method used to estimate ML models' skill and its primary purpose was to evaluate models' generalization capability on unseen data. This process had a simple parameter called k that represented the number of groups that given datasets were to be split with after random shuffles, and in this approach, k was set to 5 by default. The evaluation scores were stored in the end, epitomizing the model (Ranjan et al., 2019). The following Table 3.2 summarized the chosen parameters for each ML algorithm in GridSearchCV, and the options were decided based on the previous ML algorithms studies (Badem et al., 2019; Moldagulova and Sulaiman, 2017; Pham et al., 2017; Roberts et al., 2017; Saqlain et al., 2019; Xiao et al., 2018).

Table 3.2 Hyperparameter tunning in different algorithms

Algorithm name	**Parameters**	**Options**
LRC	solver	Liblinear
	regularization (C)	1
		10
		20
		30

RFC	n-estimators	50
		100
		150
DTC	algorithms	CART (gini impurity)
		ID3 (entropy)
	splitter	best
		random
SVC	kernel	linear
		rbf
	regularization (C)	1
		10
		20
	gamma	auto
KNN	n-neighbors (K)	7
		10
		15

Parameterization for the EVC model was conducted later separately since it combined the previously mentioned models under the best hyperparameters. The hyperparameter optimization for EVC was to select between hard voting and soft voting for its voting system. Hard voting took the most frequently model predicted results as its result. In soft voting, an individual classifier provided a probability that a specific data point matched a particular label, and the target label with the most significant sum of weighted probabilities was the final prediction (Saqlain et al., 2019).

Both hard and soft voting were adopted in the study. Finally, the model accuracy came from 5-fold cross-validation.

3.2.5 Performance evaluation and model deployment

In this approach, two methods were used to evaluate the performance of models: cross-validation and confusion matrix (f score). Shuffle split and cross-validation were applied from the Scikit-learn library because the conventional train-test split method cannot guarantee the original datasets were well mixed and may cause bias in the prediction. The confusion matrix comprised four different combinations of predicted and actual values: true positives, false positives, false negatives, and true negatives. Here, precision and recall were applied, which refers to the proportion of the relevant results and the ratio of the total number of the relevant results correctly classified, respectively.

F-score was introduced in the evaluation system to handle classification imbalance situations. The following equations were the calculations for recall, precision, and F-score (Ohsaki et al., 2017).

$$Precision = \frac{TP}{TP+FP} \tag{3.7}$$

$$Recall = \frac{TP}{TP+FN} \tag{3.8}$$

$$F\text{-}score = \frac{2*recall*precision}{recall+precision} \tag{3.9}$$

where, TP, true positive; FP, false positive; FN, false negative.

F-score ranges from 0 to 1 and reflects how accurate the model predicts. The higher F-score is, the more accurate model predicts (Deng et al., 2016). After the evaluation of both F-score and

model score, the best ML model was chosen to build a final model from the features selected datasets after PCA.

Python was the language for ML algorithms. In the end, a web-based application was further developed through Python Flask based on the developed method for more general use through APIs.

3.3 Application for oil fingerprinting analysis

The binary approach was further applied to distinguish dispersed oil for the aid of oil fingerprinting. The detailed operation procedures and results were elaborated in this section.

3.3.1 Biomarker data entry and preprocessing

The experimental data used in this study came from the work published by Song et al. (2018) which provided a detailed description of the experiment. Alaska North Slope crude oil was used in the study. Briefly, aliquot chemically dispersed oil, and non-dispersed oil samples were generated and weathered from day 1 to day 60 in a simulated marine environment. Five types of biomarkers (terpanes, steranes, triaromatic steranes or TA-steranes, monoaromatic steranes or MA-steranes, and diamantanes) were detected and analyzed to differentiate dispersed oil and non-dispersed oil. The samples were analyzed for each biomarker in 10 days intervals, through day 1 to day 60 weathering process. However, diamantanes could not be detected in long-term weathering after 30 days, while data from the other four biomarkers were retrieved through the whole weathering process (Song et al., 2019). In terms of the missing data from the experiments, medium values were filled in from the duplicates (Wei et al., 2018). For simplification of expression of peaks, Table S1 of the supplements contains information on identified peaks, including peak codes (abbreviations) and compound names (Song et al., 2019).

Originally, except that diamantanes' dataset was made of 35 samples and 9 features due to lack of detection after 30 days, the other four biomarkers' datasets were generated from 108 to 125 samples corresponded to 9 to 22 features in WCO and CDO. Quotients between each biomarker were calculated as diagnostic ratios for feature engineering. Terpanes increased to 462 features, followed by 182 features from steranes.

After feature engineering, PCA was applied for feature selection and dimensionality reduction. As mentioned in section 2, all datasets from five biomarkers were standardized before conducting PCA, and 95% variance was chosen to ensure the retention of most information from the original datasets. Scree plots were drawn for PCs against the percentage of explained variance (Murugan and Devi, 2019). The scree plots below showed the results from these biomarkers.

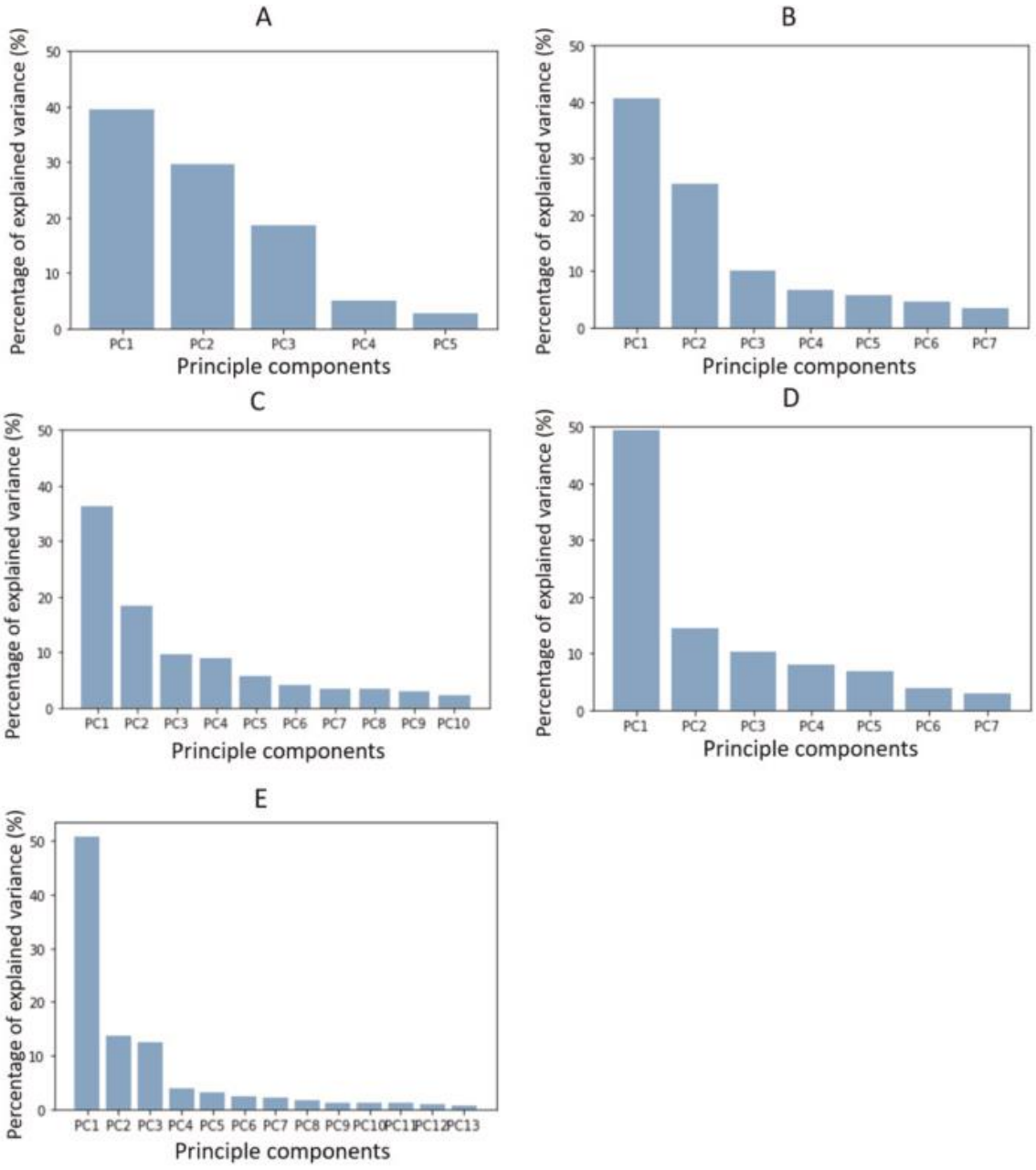


Figure 3.3 Scree plots of principal components under 95% variance of five biomarkers

(Note: A: Diamantanes, B: MA-steranes, C: Steranes, D: TA-steranes, E: Terpanes)

Comprehensively, in Fig. 3, the most PC (13) belonged to terpanes, followed by steranes (10), and the least PC belong to diamantanes, which only had 5 PCs under 95% variance. All the first PC account for above 35% variance, and the first PC in Figure. 3(D) and (E) further reached around 50% variance. Moreover, only the top three PCs of steranes in Fig. 3 (C) did not explain

more than 75% variance. Table 3.3 presented the selected features in every biomarker and features corresponded with biomarkers listed in Table S1 of the supplements.

Table 3.3 Selected features in five biomarkers

Features	Diamantanes	MA-steranes	Steranes	TA-steranes	Terpanes
PC1	9/2	1/12	16/1	3b/7	C25/H35S
PC2	5/4	4/11	14/18	3b/1	H35S/H32R
PC3	1/4	5/4	8/7	8/6	C25/C27b
PC4	5/3	12/6	16/13	6/4	H35S/H34R
PC5	6/7	11/8+9+10	17/16	6/7	C26/C28a
PC6		6/7	8/10	3b/3a	H32S/H32R
PC7		8+9+10/12	11/8	2/1	H31S/H31R
PC8			8/11		H32R/H33S
PC9			15/13		Ts/Tm
PC10			13/14		H33R/H34S
PC11					H31R/H32S
PC12					Tm/C30
PC13					C23/C25

The chemical compositions of these biomarkers decided whether they had high resistance to physicochemical or biological weathering processes (e.g., terpanes) or were more likely oxidized and biodegraded (e.g., diamantanes) (Song et al., 2019). This different response determined the number of selected features and implied potential candidates for distinguishing WCO and CDO, for example, 3,4-dimethyldiamantane, and 4-methyldiamantane in diamantanes;

C21 5ß monoaromatic steroid, and C28 monoaromatic 5α (H)-ergostane (20R) in MA-steranes; C28 20R-5α(H),14α(H),17α(H)-ergostane, and 20S-13β(H),17α(H)-diasterane in steranes; C22 triaromatic steroids (b), and C27 triaromatic-ergostane (20R) in TA-steranes; C25 tricyclic terpane (a), and 22S-17α(H),21β(H)-30,31,32,33,34-pentakishomohopane in terpanes.

For dimensionality reduction, instead of transforming features into a lower dimension, the features with the most significant loading values were chosen to represent the PCs for better interpretation and the convenience of the input for model prediction. Two more feature spaces were created for comparative analysis in terms of the impact of different PCs on the performance of ML algorithms. The original dataset was marked as X1, and the dataset with all PCs after PCA was marked as X, and the dataset with the top 3 PCs was marked as X3.

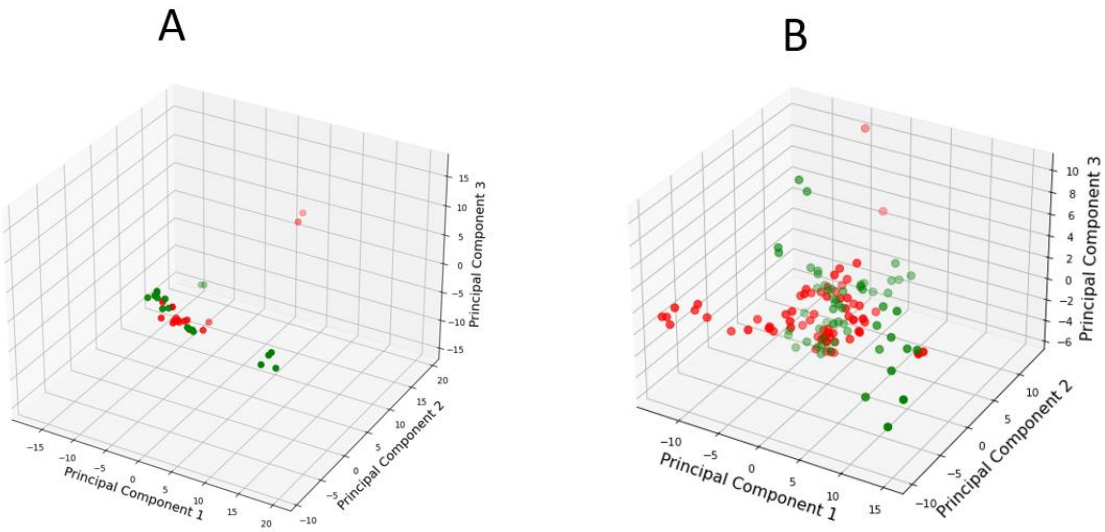
3.3.2 Data visualization

Other than resizing the datasets, dimensionality reduction also made data visualization possible by choosing the top 3 PCs by their variance. 3D graphs could be presented to generalize how data spread out in different directions. In Fig. 3.4, red dots represent CDO, and green dots represent WCO. From Fig. 3.4 (A-E), the green dots and red dots were intermixed with each other, and it was challenging to tell whether two-dimensional feature space or three-dimensional feature space separated better between WCO and CDO by visual observation. This principle could also be inferred for other dimensional feature spaces. Moreover, complicated mingle between red dots and green dots also made it difficult to tell which biomarker better separate WCO can and CDO. The interweave between CDO and WCO was caused by the addition of dispersants in oil and water emulsion because dispersants can substantially lessen the interface tension and change physical attributes, such as oil viscosity (Torres et al., 2020). These effects from dispersants could further

influence the fate and behaviours of CDO, leading to a decrease in fingerprinting accuracy (Song et al., 2018).

3.3.3 Models development for oil fingerprinting

Six ML algorithms and three different feature spaces were used for model training and comparative analysis. The cross-validation on testing datasets in GridSearchCV was set to 20%, which assigned 20% data on test datasets. The hyperparameter optimization of each model was followed by the options in Table 3.4 to find the best parameters combination. For further prediction and evaluation, the labels were converted from texts to numbers, 0 representing CDO and 1 representing WCO. The results were shown in Table 3.4 by the hyperparameter optimization.



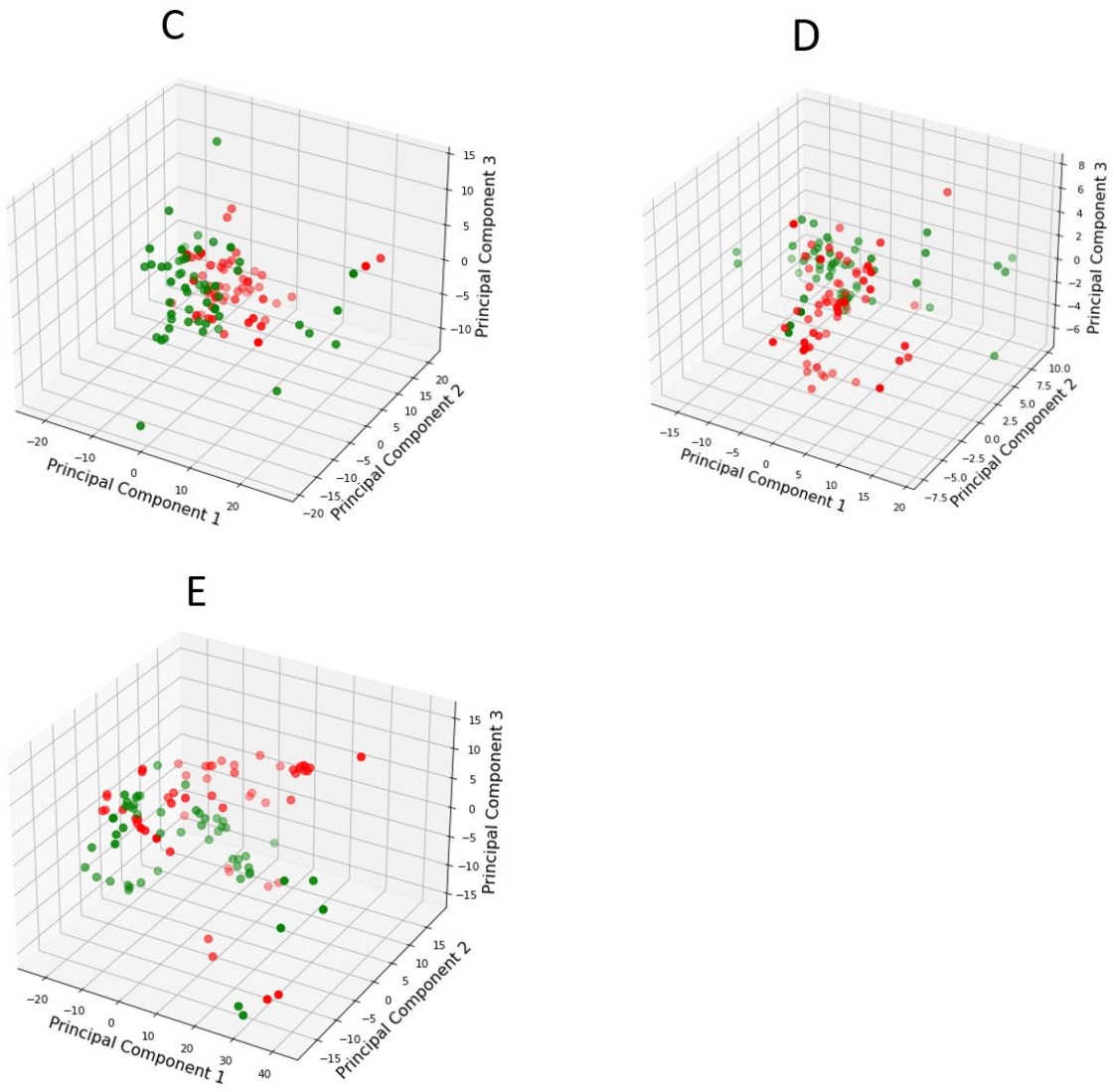


Figure 3.4 Scatter plots of the top three PCs datasets in five biomarkers

(Note: A: Diamantanes, B: MA-steranes, C: Steranes, D: TA-steranes, E: Terpanes)

Table 3.4 Outputs of different ML models in each biomarker

Biomarkers	Datasets	Models	Best model score	Best parameters
Terpanes	X (all PCs)	LRC	0.827	C: 20, solver: liblinear
		RFC	0.863	n-estimators: 100
		DTC	**0.890**	**Criterion: gini, splitter: best**
		SVC	0.842	kernel: linear
		KNN	0.727	n-neighbors: 10
		EVC	0.864	Vote: soft
TA-steranes	X (all PCs)	LRC	0.825	C: 30, solver: liblinear
		RFC	0.875	n-estimators: 100
		DTC	0.841	Criterion: entropy, splitter: random
		SVC	0.783	kernel: linear
		KNN	0.683	n-neighbors: 7
		EVC	**0.900**	**Vote: soft**

Continued Table 3.5 Outputs of different ML models in each biomarker

Biomarkers	Datasets	Models	Best model score	Best parameters
Steranes	X (all PCs)	LRC	0.727	C: 1, solver: liblinear
		RFC	**0.845**	**n-estimators: 100**
		DTC	0.800	Criterion: entropy, splitter: best
		SVC	0.736	kernel: rbf
		KNN	0.773	n-neighbors: 7
		EVC	0.800	Vote: soft
MA-steranes	X (all PCs)	LRC	0.704	C: 1, solver: liblinear
		RFC	0.848	n-estimators: 100
		DTC	0.792	Criterion: gini, splitter: random
		SVC	**0.864**	**kernel: rbf**
		KNN	0.840	n-neighbors: 5
		EVC	0.832	Vote: soft
Diamantanes	X (all PCs)	LRC	0.686	C: 1, solver: liblinear
		RFC	**0.900**	**n-estimators: 30**
		DTC	0.743	Criterion: gini, splitter: best

	SVC	0.771	kernel: rbf
	KNN	0.729	n-neighbors: 5
	EVC	0.686	Vote: soft

For each biomarker, the best-performed models were highlighted. Table S2 in the supplements provided all the parameters from three different datasets in five biomarkers. Among three datasets, the models associated with all PCs feature space after PCA acquired the highest scores. The reason could be PCA filtered out the noise in the original datasets and help improve the model prediction. Similar results of improvement of ML performance could also be found in previous PCA integrated ML studies (Nasution et al., 2018; Xu and Wang, 2005). In Table 3.4, TA-steranes and diamantanes gained the highest score (0.900) from the EVC model and RFC in the X dataset with seven features and five features, respectively. The least score (0.845) was from the RFC model in the X dataset of steranes with 10 features. These scores reflected how accurate ML algorithms predict these biomarkers.

3.3.4 Confusion matrix and decision boundary visualization

Except for model scores, confusion matrices were also introduced to probe the detailed performance of best models from a summary of prediction results, and combined with decision boundary, the intuitive understanding of how targets get divided can be easily achieved.

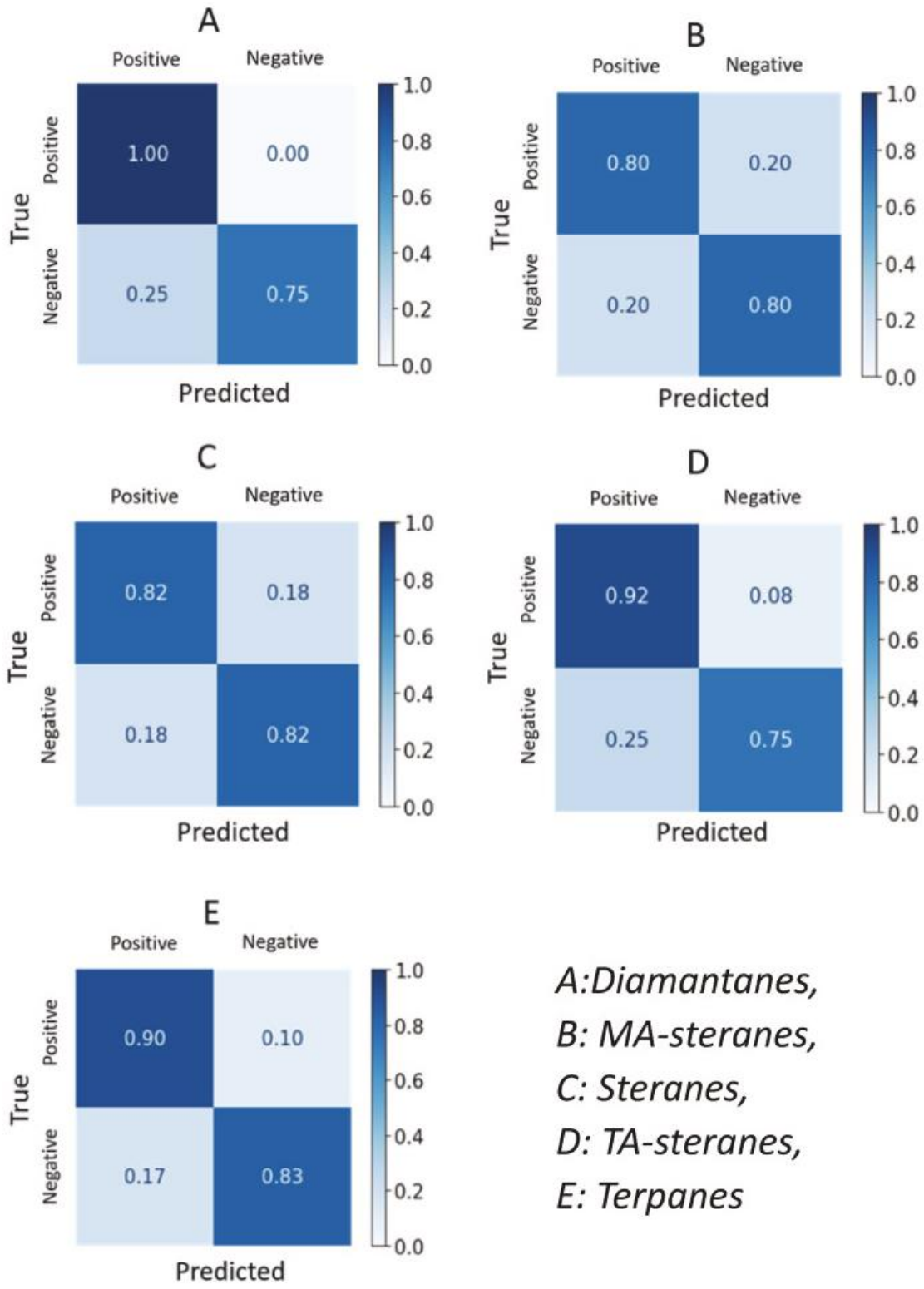


Figure 3.5 Normalization confusion matrix for the prediction of two classes (WCO and CDO)

Figure. 3.5 provided confusion matrices after training models with the best parameters and datasets with all PCs of five biomarkers from the results of Tables 3.3 and 3.4. In the confusion matrix, the small squares of top left, top right, bottom left, and bottom right represented true negative, false positive, false negative, and true positive, respectively. If evaluation only considered true negative, diamantanes Figure. 3.5 (A) had the highest rate, while terpanes Figure. 3.5 (E) had the highest true positive rate. This imbalance of precision and recall made it difficult to decide which biomarker predicted better. Hence, in this case, the f-score was set as a standard to rank models' performance on biomarkers (Juba and Le, 2019). The true values and predicted values were used to calculate the precision and recall and f-score through equations 7, 8, 9, and the final results were normalized with a color bar. F-score is a measure of a model's accuracy on a dataset, the higher F-score is, the more accurate model predicts (Ohsaki et al., 2017).

Table 3.6 F-score for different models

Biomarkers	Models	F-score
Terpanes	DTC	0.869
TA-steranes	EVC	0.792
Steranes	RFC	0.845
MA-sterane	SVC	0.793
Diamantanes	RFC	0.871

Table 3.5 listed F-scores from the best-performed models and each biomarker has different best-performed model due to the unique distribution in their datasets which was contributed by distinctive features. RFC in diamantanes performed best (0.871) among other biomarkers under the conditions of this study. The result was in consonance with our previous studies which showed

diamantanes were optimal to be used for distinguishing WCO and CDO because it was influenced by weathering (Wang et al., 2006) and dispersants also had a direct impact on it (Song et al., 2019). Decision boundary that partitioned the feature space into two labels under two dimensions. It reflected on how well different classifiers perform to separate WCO and CDO. Under the best performance in two dimensions, the targets could be separated well without missing many targets, and the results got better as dimension increases to all PCs.

Finally, ML models were deployed through the flask server and NumPy, and Pickle libraries were applied through the PyCharm for general use. The web app provided five different biomarkers options, and users could choose the one(s) for prediction to meet their requirements. After selecting specific biomarkers and filling in the input features, the result could be showed to indicate whether a sample is either WCO or CDO.

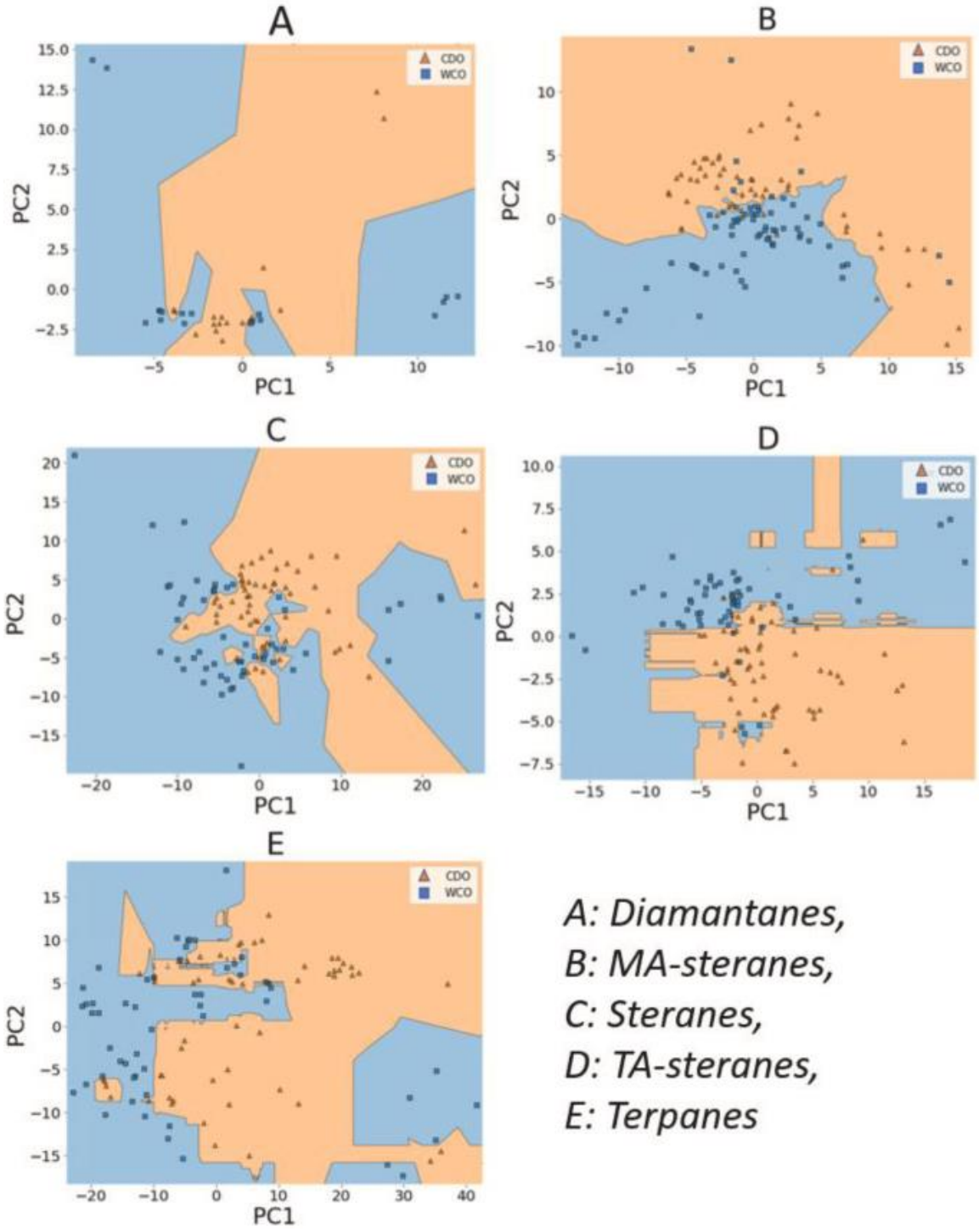


Figure 3.6 Decision boundary of models in different biomarkers

3.4 Discussions and future research perspectives

Distinguishing the dispersed oil in the marine environment is an essential step in environmental forensic. Currently, there were few of studies on applying ML methods for oil fingerprinting. This study presented a binary classification approach to classify WCO and CDO rapidly and accurately, using PCA and six ML algorithms. The performances were evaluated and compared among ML models that were trained from scratch. The criteria for choosing ML algorithms were flexible since each algorithm came with advantages and disadvantages. Hence, adequate trials of ML algorithms were made to achieve the best performance.

In this paper, PCA was implemented as a dimensionality reduction and feature selection tool to acquire a meaningful data interpretation by choosing the most significantly correlated variables. There were three benefits of dimensionality reduction. It reduced ML algorithms' training time because of the downsizing of datasets by selecting features (Becht et al., 2019). It removed noise and reduced overfilling by keeping the relevant ones, which was beneficial in oil spill fingerprinting because of the high dimensionality. Screening out the redundancy from these features could also increase the modeling accuracy (Kiarashinejad et al., 2020). It could also reflect the potential pattern or trend in datasets by choosing the top three PCs for visualization (Murugan and Devi, 2019). The following table listed the average time spent between dataset X1 (without PCA) and dataset X (with PCA) in five runs and the increasing ratio from applying X to X1.

Table 3.7 Time cost on algorithms with and without PCA

Time (s)	Diamantanes	MA-steranes	Steranes	TA-steranes	Terpanes
X1 (Without PCA)	1.929	2.240	2.437	2.230	2.872
X (With PCA)	1.826	1.862	1.920	1.925	1.913
Increasing ratio	5.6%	20.3%	26.9%	15.8%	50.1%

All biomarkers' executing time increased when training ML models with datasets before PCA and terpanes increased the most (50.1%) shown in Table 3.6. The execution speed of the dataset with PCA was faster than without PCA as it takes less time with a reduced feature set by grouping the maximum variance components in the orthogonal space (Becker et al., 2020). The amount of decreased time here might not be impressive in terms of its magnitude. However, when dealing with massive datasets, decreased training time would be more magnificent.

The sample size was used for model training in this study is relatively small (n = 35 to 125). Naturally, ML algorithms heavily rely on large datasets to increase the accuracy (Apruzzese et al., 2018). Therefore, there is a need to build an open-source database management system that gathers more information about dispersants and diagnostic ratios of biomarkers to increase dataset size and boosts the accuracy of classification in future works. Except the slow building process of database management system to boost dataset size, many other ways are also available for immediately model training. Traditional methods include synthetic minority over-sampling technique (SMOTE) (Guo et al., 2019), uniform random generation and adding noise (Moreno-

Barea et al., 2018). These methods are simple to implement but have high risk to introduce biases. The newly invented generative adversarial networks (GANs) can be substantially useful in increasing datasets and decrease the chance of introducing biases because it can simulate the distribution from sample data (Shin et al., 2018).

For applying ML models, the disadvantage is the lack of explanations that is why some ML algorithms are metaphorized as a black box because only inputs and outputs are visible and the logic behind how classification is made by features is extremely hard for a human to understand or impossible to comprehend in some cases. For example, EVC performed well on TA-steranes, but it was complex to understand how a single diagnostic ratio impacted the outcome since it involved many ML algorithms. The absence of explanation on single or local observation could cause the mistrust of models because the good performance may be contributed by noise, correlations, and overfitting (Ribeiro et al., 2016). For example, in Figure. 3.6 (D), the EVC model seems overfitting in that decision boundary. Hence, the improvement of this approach can also be achieved in interpretability and could further increase the reliability of models and the prediction accuracy. The corresponded techniques such as local interpretable model agnostic explanations (LIME), Shapley additive explanations (SHAP) will be introduced. Lately, neural networks (NN) have drawn massive attention because of their ability in classification, prediction, clustering, and associating. NN's application can be found in various industries, such as aerospace, automotive, robotics, and telecommunications (Abiodun et al., 2018). In the future, we would like to utilize different state-of-the-art NN in deep learning. For example, applying a combination of feature selection, model interpretation algorithms and NN, especially the convolutional neural network (CNN), for classification tasks to avoid overfitting and improve reliability and accuracy in oil fingerprinting.

3.5 Summary

Oil spill fingerprinting has been widely used for source identification by matching compositional parameters of both candidate spill sources and oil spill samples in the marine environment. However, the addition of dispersants on oil spills changes their chemical composition and makes it harder to determine the possible source. There have few studies on distinguishing dispersed oil by ML. The study proposed the binary classification approach for the classification to fill the gap between dispersed oil and source identification.

This study analyzed various variables to discover valuable features as future input, used data preprocessing and six ML algorithms for comparative analysis, such as RFC, SVC, KNN, LR, EVC and DT, developed the binary classification model to identify WCO and CDO. The proposed method classified sample oil as dispersed in seconds by inputting selected diagnostic ratios and prediction looked promising. The EVC and RFC models performed best with TA-steranes and diamantanes datasets in terms of accuracy, but from the decision boundary perspective, overfitting occurred with the EVC models. Thus, considering the f score and overfitting problem, RFC with diamantanes for dispersed oil classification was recommended. The preprocessed datasets with dimensionality reduction algorithms also indicated better prediction and faster speed than the original datasets. The results from this research could effectively support oil spill source identification and proved the practical value of adopting ML to facilitate oil fingerprinting.

CHAPTER 4: AN INTEGRATED APPROACH OF OPTIMIZED LEARNING NETWORKS FOR CLASSIFYING OIL-MIXED MICROPLASTICS*

*The chapter is based on a submitted article: Chen YF, Chen B, Yang M, Xin X, Kang Q, Ye XD, Zhang BY (2021) An integrated approach of optimized learning networks for classifying oil-mixed microplastics. *Journal of cleaner production*. Reference number: JCLEPRO-D-21-22184 (under review). Contributions: Yifu Chen: methodology, software, validation, paper drafting and editing; Bing Chen: conceptualization, writing, review and editing; Min Yang: resources and review; Xiaying Xin: review; Qiao Kang: data curation and review; Xudong Ye: visualization and review; and Baiyu Zhang: review and editing. All authors have read and agreed to the published version of the manuscript

4.1 Introduction

Deep learning (DL), as one of the advanced computational methods, has been widely applied in finance, medical treatment, self-driving and causality inference domains for classification and regression (Duan et al., 2020; Kang et al., 2021; Ozbayoglu et al., 2020; Sahiner et al., 2019). Among DL algorithms, convolutional neural networks (CNNs) are among the most popular networks that usually perform classification under computer vision tasks through images. The significant benefits of CNNs over other DL algorithms are that CNNs can learn features without any human intervention and weight sharing (Wu et al., 2018). Specifically, each feature map after extraction is linked to the preceding layers through a series of weights, and the locally weighted sum is then transferred to the activation layer (Qian et al., 2018). The successful application of CNNs can be seen in many computer vision domains, such as facial recognition, image search, and natural language processing (Liu et al., 2018; Sajjad et al., 2020; Shen et al., 2018).

However, CNNs also have some disadvantages, and if they are not well addressed in the application, the performance will be heavily affected. The filters can be used to check whether features are present by striding the image, and the information regarding the composition and location of the components is lost during the process (Zhu et al., 2017). Moreover, CNNs cannot classify images with different positions and understand coordinate frames of images (Tu et al., 2019). Finally, training CNNs requires a large amount of data, and collecting sufficient data may not be possible or cost-efficient due to restrictions in some domains. Therefore, some data augmentation techniques are usually introduced to solve the above difficulties. One of the most used and traditional data augmentation techniques is transforming original images by rotating, flipping, zooming in/out, and other methods. However, if data augmentation solely depends on

this technique, the model's performance may plateau or even overfit since "new" images came from the small original images' dataset (Shorten and Khoshgoftaar, 2019). Instead, deep convolutional generative adversarial networks (DCGANs) derived from generative adversarial networks (GANs) can perform better data augmentation to increase CNNs performance by increasing the size and generalization of the dataset.

Compared to the architecture of original GANs, DCGANs replaced the deterministic spatial pooling function with convolutional nets, eliminated connected layers, and included batch normalization to stabilize learning and help poor initialization (Fang et al., 2018). Its application covers from generating realistic photographs, semantic-image-photo translation to 3D object generation (Chen and Hays, 2018; Luo et al., 2020; Volokitin et al., 2020). However, there are some issues after implementing these changes in DCGANs, including mode collapse due to binary cross-entropy loss function and vanishing gradient due to the overconfidence of the discriminator (Bang and Shim, 2021). Before integrating DCGANs with CNNs to boost model performance, the above challenges of DCGANs are optimized. Furthermore, local interpretable model-agnostic explanations (LIME) are implemented to interpret the prediction and increase model reliability by highlighting salient superpixels of the input image (Zhang et al., 2018).

The mentioned classification techniques, from using DCGANs to augment and generalize datasets and applying LIME to explain the prediction from CNNs, can potentially improve the performance and reliability of CNNs models in classifying image data. To assess the performance of classification techniques, scanning electron microscopy (SEM) images of microplastics (Ostle et al.) associated agglomerates are applied since no DL-related research has been done yet. SEM has the advantage of avoiding radical illumination effects on the pixel level and background clutter

(Goldstein et al., 2017). Previous studies used SEM to discover the surface properties of the target, for example, the investigation on the surface morphology of algae cells growth (Xin et al., 2019); the polysulfide polymer in oil spill remediation (Worthington et al., 2018). However, it is rare to see the combination of SEM and CNNs in the environmental field, and their potential application related to oil spills has not been fully discovered yet.

The presence of MPs has drawn the scientific community's attention worldwide since the rising concern of its negative impacts on the ocean ecosystem and economy (Brandon et al., 2016; Kuo et al., 2018). As the main component of making surgical masks, plastic usage spikes during the COVID19 pandemic, and MPs production increases accordingly. Usually, ten percent of plastic products end up in the ocean (Magnier et al., 2019). MPs are easy to be weathered through mechanical tension (e.g., wave motion), photooxidation, and biological degradation in the marine environment to form weathered MPs (WMPs). WMPs could interact with another ocean pollutant, spilled oil (Shan et al., 2020). Spilled oil can be broken down into tiny droplets by dispersants and become much easier to be biodegraded in the marine environment (Merlin et al., 2021; Ye et al., 2021). However, when WMPs encounter the dispersed oil, they will interact with each other by forming WMPs-oil-dispersant agglomerates (WMODAs) (Yang et al., 2021).

Due to the hydrophobicity of WMPs, oil would adsorb on WMPs surface, and dispersant would then cover at the out layer, forming WMODAs in seawater. The formation of WMODAs can affect the oil droplet size and the efficiency of dispersants on spilled oil. Such impacts are determined by the weathering degree of WMODAs. For example, heavy oil droplet size increases by 124% from pristine microplastics (Ostle et al.) to 21 days WMPs; the dispersion effectiveness climbs by 44% from pristine MPs to 56 days WMPs (Yang et al., 2021). Therefore, distinguishing

the weathering degree of WMODAs is essential to define their impact better on oil spill response options such as dispersant application.

This study proposed an interpretable CNNs approach powered by optimized DCGANs and then was implemented in the oil spill field to validate its applicability. The approach mainly contained the optimized DCGANs to boost dataset size, five CNNs-based DL models to compare the performance, including customized CNNS algorithm, original and transfer learning algorithms of residual network_50 (Resnet50), Visual Geometry Group (VGG16), and LIME. The developed approach was further tested by a case study on the dataset of SEM images of WMODAs from different weathering degrees. The modeling results would help demonstrate the feasibility and robustness of the developed approach and indicate its potential in extended application in the other areas where advanced imagine analysis is required.

4.2 Materials and Methods

4.2.1 Model development

The proposed approach comprises three significant modules, as shown in Figure 4.1. The first step is image preparation and preprocessing. Two data augmentation methods, ImageDataGenerator and DCGANs, are implemented to artificially create new training data from existing data to increase the dataset size. The normalization of images is implemented to reduce the bias in the training process.

The second step is model training, in which CNNs algorithms adjust the weights of the kernel through backpropagation to decrease the loss from learning features of input pictures. A

trial-and-error strategy is applied to construct CNNs. Multiple combinations of fundamental network elements, such as 3x3 convolutional layers and max-pooling layers, are tested during the construction. Furthermore, several performance metrics are implemented to discover an optimal architecture for the network. The CNNs classifiers can learn the underlying patterns behind the labeled target during the training process by feature extraction. The feature extraction is realized by many convolutional layers performing convolution operation through filters and pooling layers performing dimensionality reduction.

The last step is to pass feature vectors to the trained classifier for the final prediction of the class. Model explanation powered by LIME is then executed to explain the predicted image by highlighting the salient superpixels and increase model reliability.

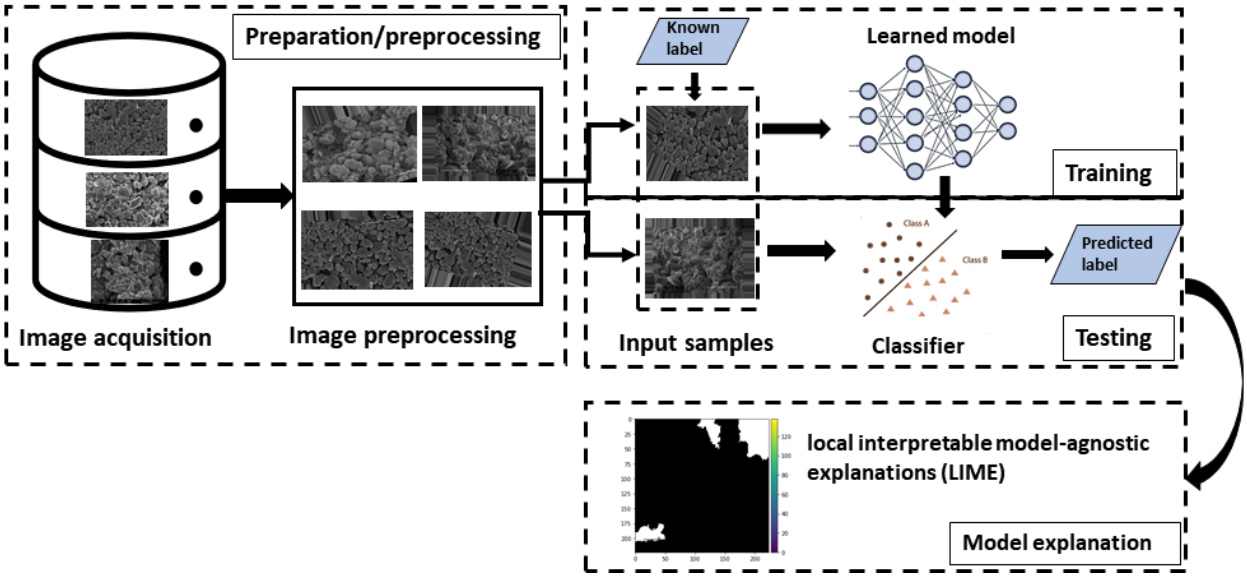


Figure 4.1 General workflow of the proposed approach for classifying WMODAs before/after 21 days

4.2.2 Optimizing DCGANs for Image Augmentation

As mentioned in the introduction, CNNs model performance can be substantially impacted by dataset size. Traditional data augmentation techniques such as ImageDataGenerator may not efficiently improve the model performance and may even weaken it since, in those techniques, "new" images are essentially simple transformations of the original images. DCGANs, as a more sophisticated data-enrichment method, can solve the problem by generating actual new images to enhance the performance. However, there have been some challenges in training DCGANs. A combination of several optimization techniques is proposed and implemented in this paper to tackle these challenges.

The first optimization shed light on weight initialization. Weight initialization is essentially helpful to accelerate DCGAN's convergence process (Dewa, 2018). From previous studies, Su et al. (2017) suggested using a zero-centered Gaussian distribution with a standard deviation of 0.02 in the dense layer. Moreover, to eliminate outliers during training, a truncated normal distribution as a variation of Gaussian distribution is applied, and it generates values from the initializer except that those values two standard deviations away from the mean are discarded. The truncated normal distribution is applied as kernel initializer in transpose convolutional layers with spectral normalization (SN) in the generator and convolutional layers with SN in the discriminator to accelerate the convergence.

As a conventional weight normalization technique, batch normalization (BN) allows networks to use a higher learning rate without compromising convergence. The benefits coming from the ability to use a larger learning rate are various. For example, the interval of the learning rate, which lies between underfitting, and gradient explosion, is more extensive. A higher learning rate also helps the optimizer avoid local minima convergence by encouraging the optimizer to

explore, and it will more easily converge on better solutions. Batch normalization transforms the input as follow:

Calculate the mean and variance of the input of the layer by equations 4.1 and 4.2:

$$\mu = \frac{1}{n}\sum Z^{(i)} \tag{4.1}$$

$$\sigma = \frac{1}{n}\Sigma\left(z^{(i)} - \mu\right) \tag{4.2}$$

Where μ means batch mean, σ means batch standard deviation, and Z means layers input.

Normalize the layers input vector $Z^{(i)}$ by equation 4.3:

$$Z_N^{(i)} = \frac{Z^{(i)} - \mu}{\sqrt{\sigma^2 - \varepsilon}} \tag{4.3}$$

Where $Z_N^{(i)}$ means the normalized Z and ε means a constant used for numerical stability.

Scale and shift to obtain the output of the layer by equation 4.4:

$$Z_O = \gamma \cdot Z_N^{(i)} + \beta \tag{4.4}$$

Where Z_O means output and γ allows to adjust the standard deviation and β allows to adjust bias.

At each iteration, the mean μ and the standard deviation σ are computed from the current batch, followed by trained γ and β through gradient descent. Even though BN benefits the performance of a network, it can still cause instability in the evaluation phase, which includes cross-validation or test and prediction. As mentioned above about the process of BN, it takes batch in the training phase to compute $\gamma, \beta, \mu, \sigma$, but in the evaluation phase, the batch is substituted by a single input at the most time. Hence, μ_pop , σ_pop are computed by using all the μ_batch , σ_batch decided in training to replace μ_batch and σ_batch in equation (Aggarwal, 2019). μ_pop , σ_pop stand for an estimated mean and estimated standard deviation of the studied population, respectively (Santurkar et al., 2018). The solution works well when the distribution of the training dataset is similar to that of the testing dataset. However, if the input came from different distribution, DCGANs might not generate ideal images since the network does not consider the actual activation values from the input. In 2018, Takeru et al. pointed out that the disjoint of model distribution and target distribution is a persistent challenge and can cause the generator to stop training because the discriminator's derivative will be 0. To stabilize the training of discriminator networks, SN is proposed, which constrains the Lipschitz constant of the convolutional filters (Zhang and Yu, 2019). Therefore, the second optimization technique is substituting BN by SN in convolutional layers in the discriminator.

Neural network training is sensitive to the minimized loss, and by performing gradient descent to minimize standard cross-entropy, networks can achieve better classification performance and fast convergence. However, more novel techniques could outperform the standard cross-entropy, for example, label smoothing. Label smoothing improves accuracy by computing cross-entropy with a weighted combination of targets from the datasets with the uniform distribution instead of the hard targets. Müller et al. (2019) experimented with how label

smoothing can implicitly calibrate a model's prediction and significantly impact model interpretability. Furthermore, they also revealed that label smoothing would impair distillation which is when teacher models are trained with label smoothing, student models perform worse. Since the application of label smoothing in DCGANs does not involve these two side effects, it is applied as the third optimization technique to calculate loss and accelerate the convergence of generator and discriminator.

The last added optimization technique is instance noise. It aims to make the true and predicted distributions overlap that help DCGAN's instability. Sequentially, fitting a custom distribution of generated images would be easier in the learning process (Mescheder et al., 2018). The application of instance noise is by adding noisy labels to calculate the discriminator's loss to help it better convergence.

Before training DCGANs, input images need to be normalized to the pixel range of -1 to 1 since the Tanh activation function is suggested in the generator for its output layer (Chen and Wang, 2020). The architecture of the generator is modified as below:

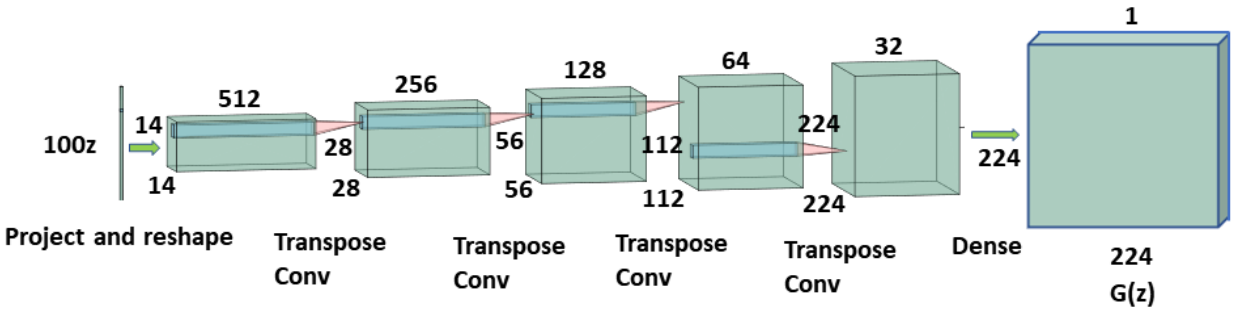


Figure 4.2 DCGAN generator used for WMODAs modeling. A 100-dimensional truncated normal distribution Z is projected to a small spatial extent convolutional representation with many feature maps.

Figure. 4.2 illustrates how the generator takes 100×1 noise vector and upsamples it by four transposed convolutional layers using 512, 256, 128, 64, and 32 different filters with a stride of 2 to the output, which is 224×224×1 to match with the input size. Discriminator reverses this process which downsamples it by using four convolutional layers with 64, 128, 256, and 512 filters each and outputs 1 and 0, which are classification classes for the real or fake image.

4.2.3 Transfer learning

Transfer learning is a machine learning methodology in which a model built for one task is used as the foundation for another task to optimize the progress and improve model performance (Zhuang et al., 2020). Two original deep neural networks, Resnet50 and VGG16, are applied for comparative study. Resnet50, as the modern network architecture representative, introduces residual blocks. It provides the alternative shortcut through identifying functions for the gradient to flow through to avoid vanishing gradient and performance degradation. The residual block is presented in Figure 4.3.

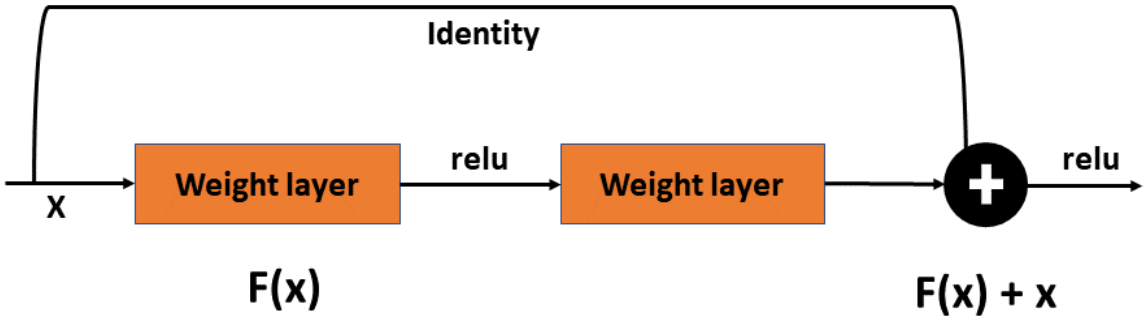


Figure 4.3 Residual block in ResNet50.

The architecture of ResNet50 is made of 50 layers, including one input layer, five convolutional blocks with ReLU activation function, and one dense layer with a SoftMax activation function (Theckedath and Sedamkar, 2020).

The VGG16 represents a classic deep neural network. Its architecture consists of 16 3x3 convolutional layers stacking on top of each other in increasing depth and using only max pooling to reduce volume size. A SoftMax classifier then follows two fully connected layers with 4,096 nodes (Qassim et al., 2018).

The pre-trained ResNet50 and VGG16 models are introduced from the Keras library, and they are trained with the ImageNet dataset, which contains more than 1.2 million images in 1000 classes.

4.2.4 Comparative study of applied CNNs models

The comparison mainly explores the performance and architecture between original models and transferring learning models. Four criteria are selected: accuracy, training time, total parameters, and model size. Accuracy indicates the correct prediction of each model, while training time refers to the processing speed of each model. Total parameters and model size specify the complexity and amount of memory required to deploy the final prediction architecture.

A normalized confusion matrix (Table 4.1) is introduced to calculate the model accuracy. The normalized confusion matrix comprises four different combinations of predicted and actual values: true positives, false positives, false negatives, and true negatives. Precision and recall from the confusion matrix are applied, referring to the proportion of the relevant results and the ratio of the total number of the relevant results correctly classified, respectively.

Table 4.1 An illustration of the confusion matrix

	Predicted positive	**Predicted negative**
Actual positive	True Positives (TPs)	False Positives (FPs)
Actual negative	False Negatives (FNs)	True Negatives (TNs)

The evaluation system also introduces F-score from the confusion matrix to handle classification imbalance situations between precision and recall. The following equations are the recall, precision, and F-score (Chicco and Jurman, 2020).

$$\text{Precision} = \frac{TP}{TP+FP} \tag{4.5}$$

$$Recall = \frac{TP}{TP+FN} \tag{4.6}$$

$$\text{F-score} = \frac{2*recall*precision}{recall+precision} \tag{4.7}$$

where TP, true positive; FP, false positive; FN, false negative.

F-score ranges from 0 to 1 and reflects how accurate the model predicts. The higher F-score is, the more accurate model predicts (Chen et al., 2021).

4.2.5 Model interpretation

The lack of explanation of how the DL model predicts has been constantly criticized. Even though CNNs can visualize the network architecture and the output of hidden layers, the feature maps and filters are still quite abstract to understand. Instead, LIME is supplemented to explain how models predict. LIME is an innovative explanation approach that leans an interpretable model locally around the prediction to explain any classifier's prediction in an interpretable and faithful manner (Visani et al., 2020). There are four significant steps in LIME: 1) create perturbation of

image through superpixels, 2) predict classes of new images by CNNs models, 3) compute distances between the original image and each of the perturbed image and computed weights of each perturbed image, 4) use perturbations, predictions, and weights to fit an explainable model (Garreau and Luxburg, 2020). The explainable model highlights the top features in superpixels which are reasons how models predict.

The network is built by the Keras (2.4.0) and TensorFlow (2.5.0) libraries, which are python libraries developed by google for deep learning applications. All techniques mentioned above are executed in Python v3.8 and run on a GTX 2060 super GPU with CUDA 11.3.

4.2.6 Case study in WMODAs images

Before making WMODAs samples, Polyethylene (PE) (6.00-8.50 um) was obtained from Micro Powders Inc. (New York, USA). Heavy oil (API gravity 8.00-15.00) was applied in this study. Corexit 9500A was obtained from Nalco Environmental Solutions LLC (Texas, USA). Sea salt was purchased from Millipore Sigma (Ontario, Canada). Ultrapure water was used throughout the experiments. Synthetic seawater was made by dissolving 34.00 sea salt in 1L ultrapure water and filter through a 0.20 μm membrane to remove suspended particles. Based on the method developed by Yang et al. (2021), 100 μL heavy oil and 4 μL Corexit 9500A were released into 120 mL synthetic seawater, reaching a dispersant to oil volume ratio (Roberts et al.) of 1:25 for the oil-dispersant-seawater mixture. To produce different WMODAs, 48mg of 7, 14, 28, 42, or 56 days weathered MPs, prepared following Yang et al. (2021), were added to the mixture. Then the mixture was shaken for 10 min at 200 rpm and kept stationary for 10 min in each run. WMODA samples were collected and dried until no weight loss. The surface morphology of WMODAs was

observed by a FEI Quanta 650 FEG scanning electron microscopy (SEM) (Thermo Fisher Scientific, Hillsboro, OR, USA). All samples were coated with gold to improve conductivity. SEM images were acquired from five WMODAs samples, and for each class, 100 SEM images were taken to guarantee the balance between the two classes. Since CNNs cannot adapt to scaled images, ImageDataGenerator was applied to add more variance in the dataset, including horizontal flip, zoom in/out, width and height shift, rescale, and rotation.

4.3 Results and Discussion

4.3.1 Image augmentation

The optimization techniques mentioned in the methodology were implemented in the DCGANs to boost the WMODAs image dataset size. Spectral normalization was performed via power iteration operation on the specific weight for convolutional layers in the discriminator. Label smoothing set the class labels in the range [0, 0.3] for the WMODAs before 21 days and [0.7, 1] for the WMODAs after 21 days. Noisy labels were determined by adding 5% of the error to the labels.

Except for the original SEM images as the input for DCGANs, additional images from ImageDataGenerator were also included to add more variance to generate synthetic images (Shorten and Khoshgoftaar, 2019). The image dataset was trained in optimized DCGANs in 1,000 epochs and saved in the same directory as picture format. Each epoch costs about 16 seconds, and this result could be varied based on different CPUs and GPUs of the computer. In total, training WMODAs images before/after 21 days cost about 4.4 hours, and the WMODAs datasets before/after 21 days were augmented to 1,200 images and 1,200 images, respectively.

In Figure 4, the losses from the generator and discriminator were presented. In Figure 4(a), the losses from the generator, which was illustrated as the blue line, fluctuated as iterations increased, while the losses from the discriminator were relatively stable. However, as epochs increased, the losses from the generator and discriminator gradually converged, as shown in Figure 4.4(b). This trend proved that the optimizations were successful in stabilizing DCGANs' training process. The results were also reflected in the generated images of WMODAs before 21 days in Figure 4.5. In Figure 4.5(a), the nine images were generated at the first epoch, and apparently, the quality was not acceptable with coarse texture. As training went further, the image quality improved drastically. In Figure 4.5(b), the images were generated at the final epoch, and it was clear to see the details of generated agglomerates.

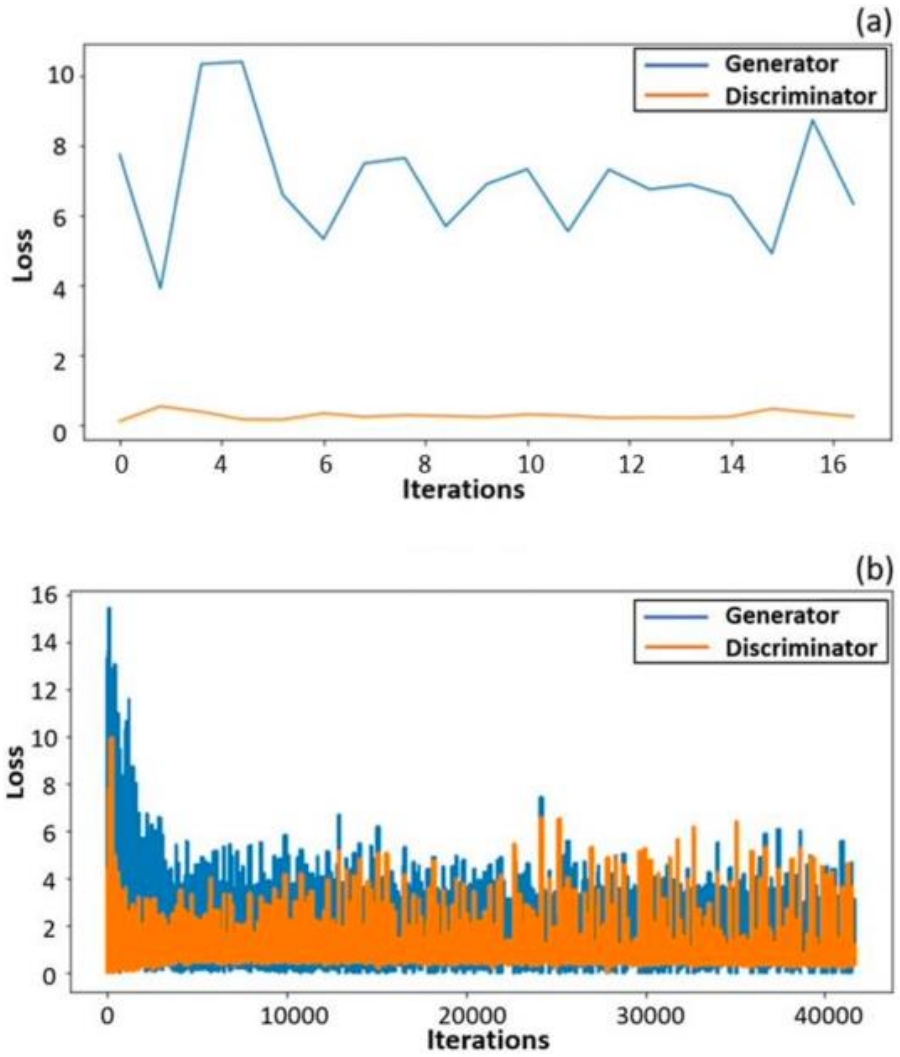


Figure 4.4 Generator and discriminator loss at epoch 1 (a) and their loss at final epoch in all iterations (b). The blue line is the generator, and the yellow line is the discriminator

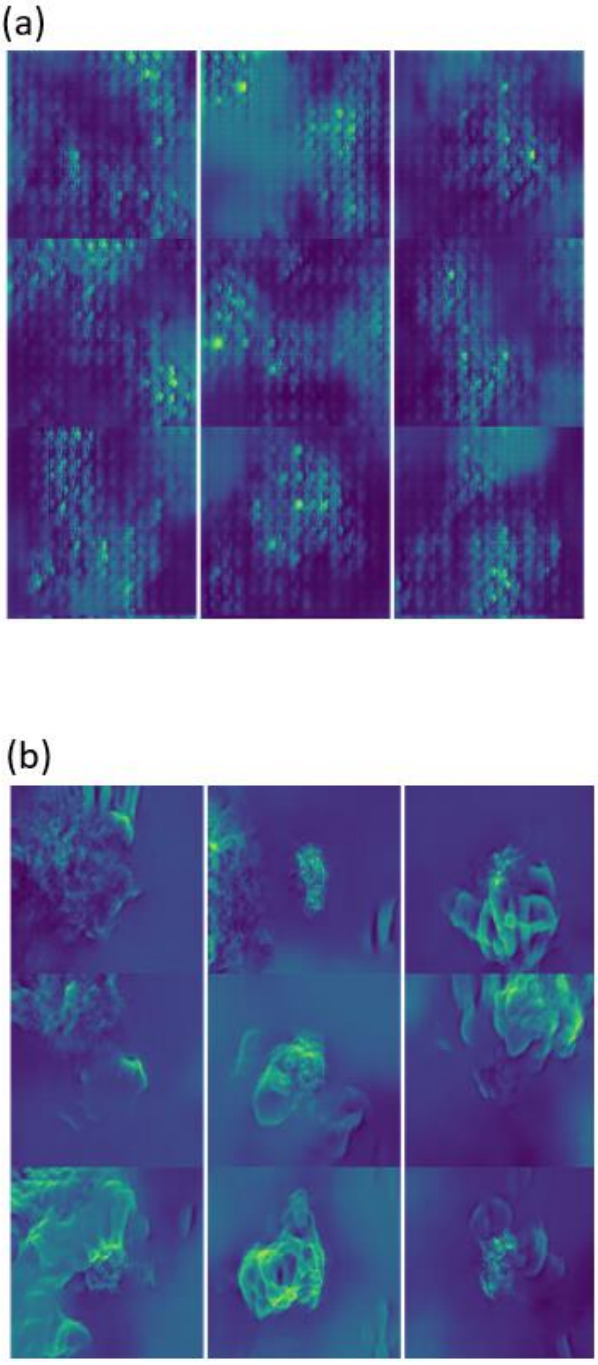


Figure 4.5 Generated WMODA images before 21-day weathering at epoch 1 (a) and the final epoch 1,000 (b).

4.3.2 Customized CNNs model (benchmark)

The augmented SEM images were the input for CNNs models. However, the input layer of CNNs only took images with a specific resolution; thus, resizing needs to be done before feeding images to CNNs since these SEM images had different resolutions. All images were resized into 224×224 pixels, and the channel was set to one since the input SEM images were grayscale. The two-dimensional network was built from scratch as a benchmark by utilizing several various convolutional and pooling layers. The convolution layers were composed of 3 × 3 filter sizes with filters from 16 to 64. Three convolution layers were passed through a ReLU function to allow a non-linear transformation. Three max-pooling layers had a filter with a size of two or five. The final output feature maps were flattened and attached to the dense layer and output layer. The 2D network architecture is shown in Figure 6.

The network's weights were initialized using the Kaiming initialization or He initialization through kernel initializer in the Conv2D layer, which is an initialization method for non-linearity of activation functions such as ReLU activation. This initialization facilitated model convergence (Sai and Lee, 2018). Following that, these weights were modified through training processes. The customized CNNs were then trained with a batch size of 5 and a maximum epoch of 1000. Furthermore, the network was programmed with early stopping on the validation set with an initial learning rate of 0.001 and was used sparse-categorical-cross-entropy for loss function and Adam for optimizer with a minimal change rate of 0.0001 and patience of five epochs.

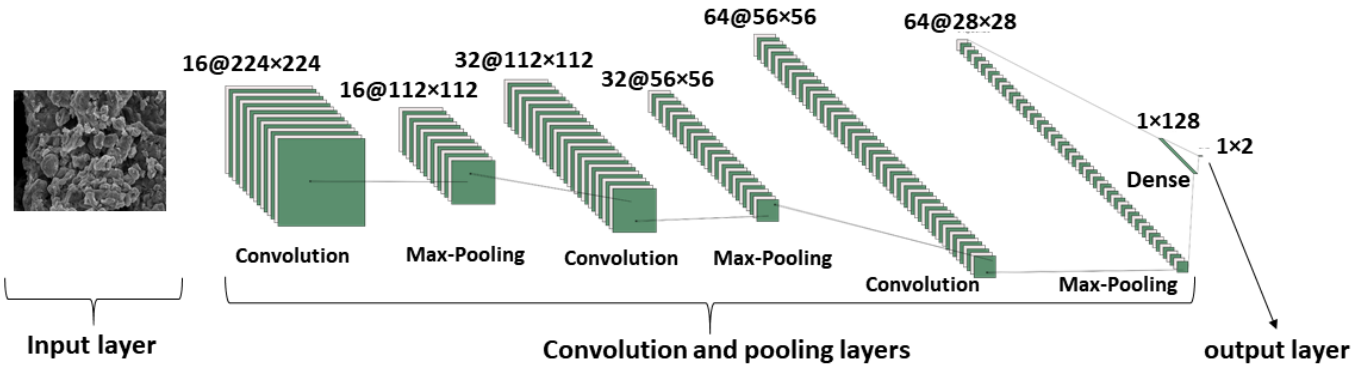


Figure 4.6 representation of the two-dimensional CNNs architecture with SEM image inputs connected to the convolution, pooling, and output layers.

4.3.3 Adaptive transfer learning

All grayscale WMODAs images were transformed into RGB channels to be aligned with the input channels of pre-trained CNNs models. Furthermore, the last fully connected layer was randomly initialized and freshly trained to accommodate the new object categories in our WMODAs study. The learning rate was kept at the default 0.01. Hence, the architectures of the pre-trained ResNet50 and VGG16 were identical to the original ones, except for the last two layers. Pre-trained weights were loaded to improve training speed and model accuracy.

4.3.4 Training results and comparative analysis of different CNNs models

In this study, we explored, evaluated, and analyzed the influence of various CNNs algorithms, including original and transfer learning algorithms from non-microplastic to microplastic image domains. Five CNNs-based models were trained with early stopping by

augmented datasets stratified into train and test subsets in the ratio of 0.8/0.2. The results of each model are presented in Table 4.2.

Table 4.2 Experimental results of different CNNs models over MP images datasets

Model	**Accuracy**	**Training time (sec)**	**Total parameters**	**Model size (Mirnaghi et al.)**
Customized CNNs model	0.6122	15	123,650	1,499
VGG16 (WOT)	0.6282	66	21,202,883	248,618
ResNet50 (WOT)	0.8738	141	23,794,578	96,170
VGG16 (WT)	0.9593	42	21,203,778	133,598
ResNet50 (WT)	0.9986	117	23,788,418	94,994

Note “WOT”: without transfer learning; “WT”: with transfer learning.

In Table 4.2, the customized CNNs model, as a benchmark model, performed the worst (0.6122), and ResNet50 (WT) performed the best (0.9986) in terms of model accuracy. All models with transfer learning acquired higher accuracy than those which were only trained with original WMODAs datasets. ResNet50 (WOT) spent the longest time training and had the largest total parameters. Noticeably, ResNet50 had more parameters in deep architecture, but their model size is substantially less than VGG16. ResNet50 applied global average pooling rather than fully connected layers in VGG16 to reduce model size. These benefits from ResNet50 can facilitate deployment in a quicker and more efficient manner, compared to other larger models.

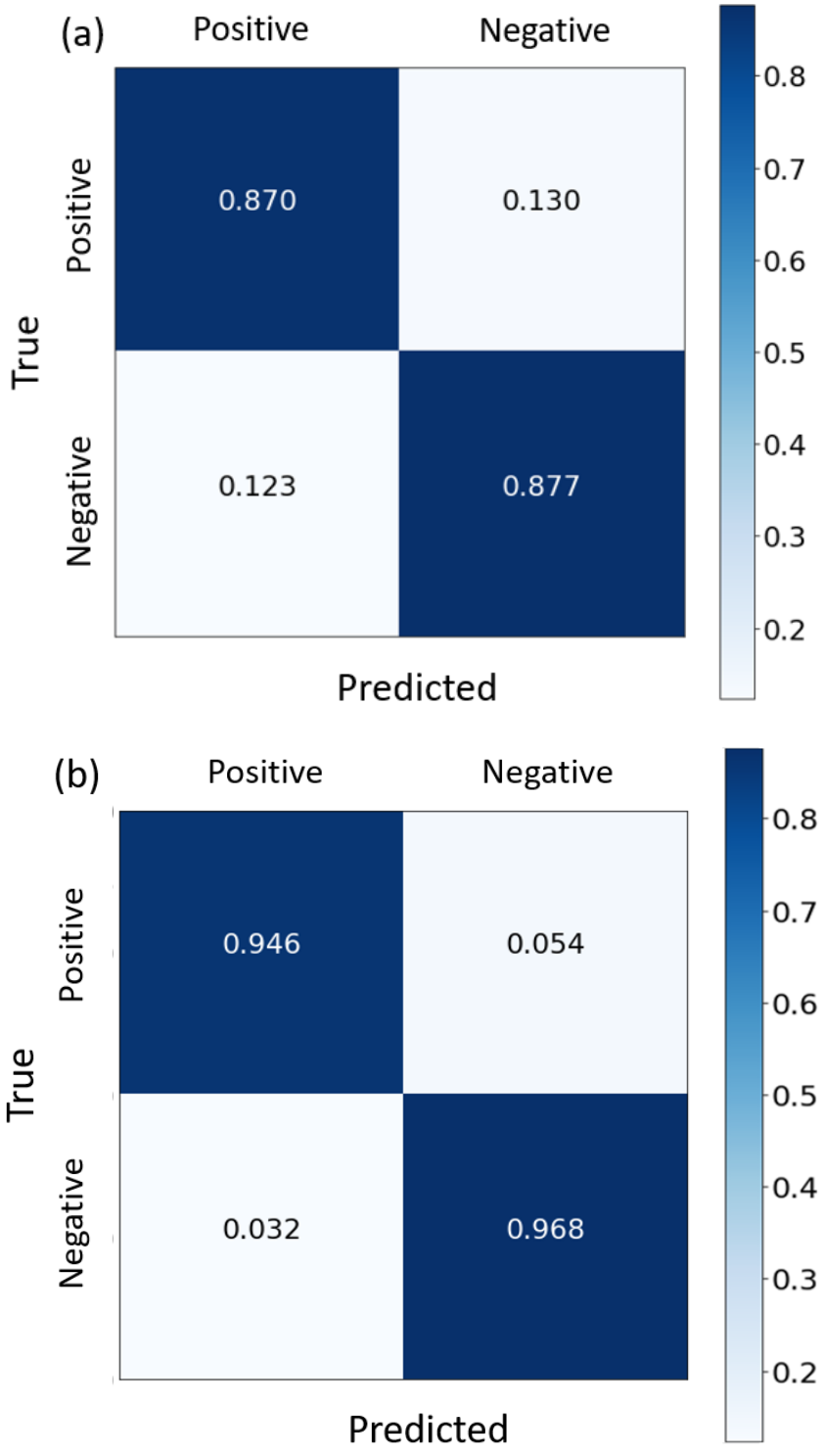


Figure 4.7 Confusion matrix of VGG16 (WT) in (a) and ResNet50 (WT) in (b).

In Figure 4.7, confusion matrices with normalization were presented from the two highest accuracies scored models in table 1. Through the equations (5), (6) and (7), F-score from VGG 16 (WT) and ResNet50 (WT) were 0.9063 and 0.9192, respectively. Hence, ResNet50 performed better with WMODAs dataset. Despite the disparity between natural images and WMODA images, CNNs were comprehensively trained on the large scale well-annotated ImageNet could still be effective when they were transferred to perform WMODA image recognition tasks.

4.3.5 Model interpretation on WMODAs image

In image preprocessing, segmentation usually played a vital role in dividing the image into segments, and algorithms would train only important segments to improve the model accuracy. However, since the SEM images were included by one object in this study, segmentation would not make much difference in model accuracy. Instead, image segmentation was combined with LIME for improving model explanation and reliability.

An increasing model explanation was essential for the reliability of applying the ML model. LIME was applied to determine the aspects of the models that were salient in regions to help recognize particular parts of images. In Figure 4.8, the primary process of how LIME makes predictions is illustrated. In Figure 4.8(a), the image which needed to be explained was extracted superpixels and divided by yellow lines. Then, through the perturbation function, the given image was perturbed based on a perturbation vector and predefined superpixels, and one of the examples was shown in Figure 4.8(b). The kernel function is by computing the cosine distance between each randomly generated perturbation and the given image. Once the weighted model

was created, the given image could be explained by choosing the top important superpixels, as shown in Figure 4.8(c).

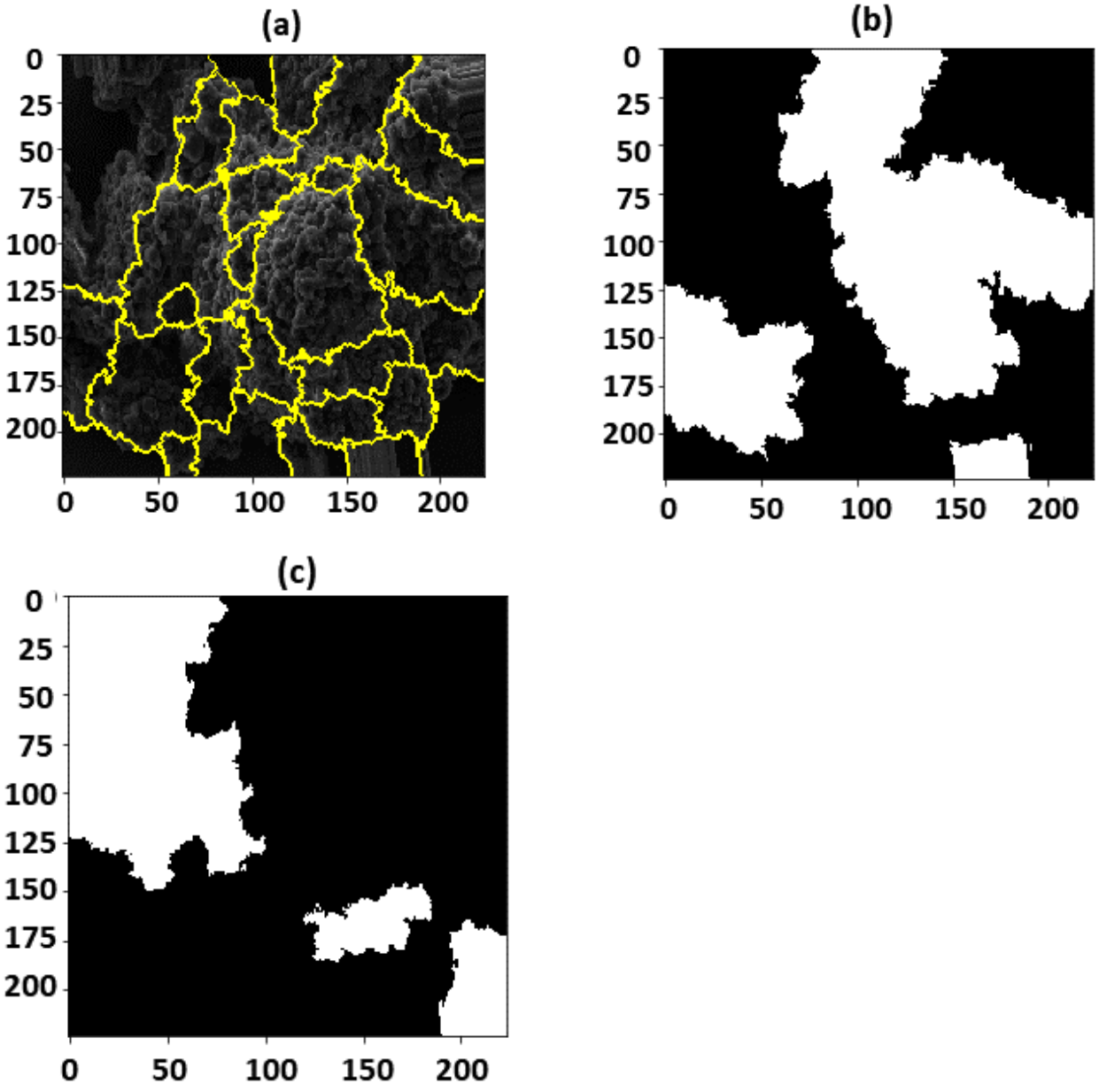


Figure 4.8 Explanation of a prediction of WMODAs before 21-day weathering with LIME: (a) interpretable components (adjacent superpixels), (b) an example of a perturbated image, and (c) an explanation image

4.4 Summary

The effects on dispersants can vary from different weathering degrees of WMODAs and currently, there has no effective method to distinguish them. The deep neural network is a promising method in image classification and can potentially aid the problem of separating WMODAs from different weathering degrees. However, the application of deep neural networks always faces a dilemma: users fear overfitting because of data shortage. Therefore, it is not widely implemented in fields that usually do not generate substantial data.

In this study, an interpretable CNN approach with optimized DCGANs was developed and successfully applied on the classification task of distinguishing WMODAs before/after 21 days using SEM images. To our best knowledge, this was the first time that proposed image classification techniques have been trained to categorize WMODAs in the marine environment.

A combination of several optimization methods was applied to effectively stabilize DCGANs' training process to generate high-quality images, including truncated normal distribution, spectral normalization, label smoothing, and instance noise. The F score of 0.9192 and the accuracy of 0.9986 was achieved by pre-trained Resnet50 with augmented datasets, indicating a robust prediction result. LIME was applied to explain the developed model in predicting WMODAs based on SEM images, which successfully increased the model's reliability.

The developed approach can have positive impacts on marine and modeling studies based on SEM images. From a marine environmental aspect, the developed approach can assist the understanding of the effects by differentiating weathering degrees of WMODAs. The formation of WMODAs would affect the efficiency of oil spill response options such as dispersant

application in oceans. Furthermore, in WMODAs, the MP weathering degree played an important role in oil dispersion effectiveness. As the weathering process helped form hydrophilic functional groups (OH, C=O, COC—) in WMPs, the increase of MP hydrophilicity led to less assumption of dispersant in WMPs dispersion. As a result, more dispersants would be available for oil dispersion. Thus, classifying weathering degree of MPs would be crucial for understanding their effects on oil dispersion effectiveness in offshore oil spills.

Furthermore, the application of the developed approach could also help characterize other WMPs associated with agglomerates and be implemented in classifying SEM images in other fields, such as environment, materials, and broader engineering applications.

CHAPTER 5: CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusions

In this dissertation research, the history of machine learning (ML) development was first briefly introduced, and more details of standard ML algorithms were provided in supervised and unsupervised learning, such as SVM, KNN, logistic regression, PCA, and others. Data preprocessing is one of the necessary steps to conduct before training the model to increase the model accuracy. Four different strategies in data preprocessing were elaborated, including data cleaning, data integration, data transformation, and data reduction. Data augmentation techniques were necessary before training the model when dealing with small datasets. Mainly, when training deep neural networks, large datasets were better to free the model from the limitation of small datasets, and data augmentation techniques could boost the size of datasets when large datasets were not available. Two significant data augmentation techniques, geometric transformation, and GANs were introduced to augment image data. Following up the tremendous potential in tackling data patterns, several applications of ML in environmental fields were showcased, but there have been few studies of ML targeting the problems related to oil fingerprinting in an oil spill and MPs in marine pollution. Therefore, these two fields' problems were explained in Sections 2 and 3 under Chapter 2, and the approaches were proposed in Chapters 3 and 4.

In Chapter 3, the proposed pipeline targeted sequential data. It is integrated with data preprocessing, feature engineering, hyperparameters tunning, and deployment of the final model. Five popular ML algorithms and one ensemble ML algorithm based on the weights of the first five ML algorithms were trained and evaluated. It then applied to the problem of WCO and CDO in oil fingerprinting. The dataset of diagnostic ratios from five different biomarkers was extracted to perform a training job from a previous study that the oil sample provided by Alaska North Slope.

After conducting PCA for dimensionality reduction and visualization in data preprocessing, the preprocessed dataset was passed to hyperparameter optimization. The results involved two comparisons between different ML algorithms and different biomarkers, respectively. After the evaluation from the confusion matrix, the best F score (0.871) was achieved by RFC from the diamantanes biomarker. The final model was then deployed through flask API on the website for public access. The proposed pipeline can handle tabular datasets in many other environmental fields, such as water, air, or soil. A slight change can be made by selecting different ML algorithms considering the characteristics and distribution of specific datasets. The developed ML model has proved the successful application in predicting WCO and CDO and can support oil spill source identification in environmental forensic effectively.

In Chapter 4, to cover more comprehensive ML applications in the environmental domain, classifying image data through integrated neural networks was proposed. The proposed approach for image classification included image preprocessing (data augmentation), constructing the CNN model, and model explanation. The major problem in many ML training was data shortage, and this problem was magnified in many environmental domains, especially in image data format. Hence, the approach adapted optimized DCGANs for not only grayscale but RGB images augmentation as well. Through introducing SN, label smoothing, and instance noise, DCGANs training became more stable and generated better images. The augmented image dataset was then applied to train different CNN models, including transfer learning models and original models. LIME was implemented to explain the prediction and increase the reliability of the developed model.

Meanwhile, the explanation from LIME can also improve the model accuracy by adjusting the dataset accordingly. The proposed approach was then applied to classify WMODAs with

different weathering degrees in marine pollution. The grayscale SEM image dataset of WMODAs was selected to perform classification. From the comparison of F score, ResNet50 with transfer learning reached the highest score (0.9192), and it had a smaller model size than VGG16 due to its residual block design. With the image analysis of the developed model, distinguishing WMODAs from different weathering degrees became easier and provided a fundamental understanding of how they can affect oil dispersion effectiveness in offshore oil spills.

From the two case studies illustrated in this research, the characteristics of ML and the situation of choosing ML in the environmental field can be summarized. In the first case study, the dataset of oil fingerprinting potentially contained hundreds of features, and previous studies only chose a dozen diagnostic ratios as features by heuristics because traditional data analysis cannot handle hundreds of features simultaneously. Moreover, using the trained ML model to predict oil type was not applicable by inputting hundreds of features. Therefore, PCA was supplemented to conduct dimensionality reduction to simplify the input process while potentially increasing prediction accuracy by eliminating high correlated features and constructing new features. Lastly, in terms of increasing the mobility of the application, the final model should be deployed on a website or mobile app. By considering the above reasons, the ML was decided to deal with the oil fingerprinting problem. The second case study was a computer vision-related environmental problem, and traditional analysis could not deal with this type of data. The significant contribution was integrated optimized DCGANs to boost the dataset and LIME to explain the prediction. GANs training was difficult and faced many challenges, such as mode collapse, convergence problem, etc. Many variations of GANs were proposed, but there was no perfect GANs to be trained to generate any datasets. Due to convolutional layers' ability in feature extraction, DCGANs were chosen to perform augmentation on the WMODAs image dataset. To further free CNN models

from the limitation of dataset size, transfer learning models trained on hundreds of thousands of images were selected to perform classification. It was the first time GANs' variation coupled with LIME as image analysis tools to perform classification in the WMODAs field. The two case studies environmental domains demonstrate the success of proposed ML approaches in dealing with sequential and image datasets, and they also showed how to design ML approaches to address different environmental challenges.

Based on the research, three journal papers and conference presentations are under review or preparation as follows:

Yifu Chen, Bing Chen, Min Yang, Xiaying Xin, Qiao Kang, Xudong Ye, Baiyu Zhang. An integrated approach of optimized learning networks for classifying oil-mixed microplastics. (Submitted, reference number: JCLEPRO-D-21-22184). The Journal of Cleaner Production. My duty is developing the proposed approach, building the case study model, analyzing results, and writing the whole paper.

Yifu Chen, Bing Chen, Xing Song, Qiao Kang, Xudong Ye, Baiyu Zhang. A data-driven binary-classification approach for oil fingerprinting analysis (published). Environmental Research. DOI: https://doi.org/10.1016/j.envres.2021.111454. My duty is developing the proposed approach, building the case study model, analyzing results and writing the whole paper.

Min Yang, Baiyu Zhang, Yifu Chen, Xiaying Xin, Keneth Lee, Bing Chen. Imapct of Microplastic on Oil Dispersion Efficiency in the Marine Environment. (published). Sustainability. DOI: https://doi.org/10.3390/su132413752. My duty is conducting data analysis using machine learning.

Conference oral presentations:

Yifu Chen, Bing Chen, Min Yang, Xiaying Xin, Qiao Kang, Xudong Ye, Baiyu Zhang. Convolutional neural network approach for classifying oil-mixed microplastics. The Canadian association on water quality. The LEADERS & PEOPLE 2021 Virtual Symposium 2021, July 20-22, Virtual. My duty is making and delivering presentation.

Yifu Chen, Bing Chen, Xing Song, Qiao Kang, Xudong Ye, Baiyu Zhang. Machine learning aided classification for oil fingerprinting analysis. The Canadian Society for Civil Engineering (CSCE) 2021 Annual General Meeting and the 18th International Environmental Specialty Conference, May 26-29, Virtual. My duty is developing the proposed method and presenting at CSCE.

Yang M*, Zhang BY, Chen YF*, Chen B (2021) Effects of microplastics on oil droplet size distribution in the marine environment. The LEADERS & PEOPLE 2021 Virtual Symposium 2021, July 20-22, Virtual. My duty is analyzing image data of microplastics.

5.2 Recommendations for Future Work

In this research, a new data-driven binary classification method including dimensionality reduction and seven classification algorithms was first proposed to analyze diagnostic ratios as tabular values. The sample size for oil fingerprinting was somewhat limited. Naturally, ML algorithms relied substantially on multiple datasets to improve accuracy. The GANs based data augmentation was supplemented in the developed method for image analysis. However, in the approach of analyzing tabular values, it was only augmented the traditional geometric transformation without bringing more variations into the dataset. In the future, other data augmentation techniques, such as autoencoder and GANs with pre-trained generators, can be

tested on biomarkers' datasets and choosing the best-performed technique based on the distribution of the specific dataset.

The explanatory algorithm was applied in the developed image analysis method; however, it was difficult to tell the region of interest (Roberts et al., 2017) without the help of an expert in MPs since the difference between each weathering degree is very subtle. In the future, the dataset with the highlights of the region of interest (ROI) could be established with the support of experts in MPs to help identify the highlighted areas in prediction coming from the explanatory algorithm.

In addition, other novel methods may be considered in future research. For example, you only look once (YOLO) algorithm has the potential for the prediction in real-time and can be used in environmental data analysis to detect any changes in shape, color, and other properties of targeted elements.

REFERENCES

Aggarwal, s. L. P., 2019. Data augmentation in dermatology image recognition using machine learning. Skin Research and Technology. 25, 815-820.

Anderson, P. J., Warrack, S., Langen, V., Challis, J. K., Hanson, M. L., Rennie, M. D., 2017. Microplastic contamination in lake Winnipeg, Canada. Environmental Pollution. 225, 223-231.

Ayodele, T. O., 2010. Types of machine learning algorithms. New advances in machine learning. 3, 19-48.

Abdi, H., Williams, L. J., 2010. Principal component analysis. Wiley interdisciplinary reviews: computational statistics. 2, 433-459.

Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Mohamed, N. A., Arshad, H., 2018. State-of-the-art in artificial neural network applications: A survey. Heliyon. 4, e00938.

Alizadeh, M. J., Kavianpour, M. R., Danesh, M., Adolf, J., Shamshirband, S., Chau, K.-W., 2018. Effect of river flow on the quality of estuarine and coastal waters using machine learning models. Engineering Applications of Computational Fluid Mechanics. 12, 810-823.

Ao, Y., Li, H., Zhu, L., Ali, S., Yang, Z., 2019. The linear random forest algorithm and its advantages in machine learning assisted logging regression modeling. Journal of Petroleum Science and Engineering. 174, 776-789.

Apruzzese, G., Colajanni, M., Ferretti, L., Guido, A., Marchetti, M., On the effectiveness of machine and deep learning for cyber security. 2018 10th International Conference on Cyber Conflict (CyCon), 2018, pp. 371-390.

Bach, S. H., He, B., Ratner, A., Ré, C., Learning the structure of generative models without labeled data. International Conference on Machine Learning. PMLR, 2017, pp. 273-282.

Bælum, J., Borglin, S., Chakraborty, R., Fortney, J. L., Lamendella, R., Mason, O. U., Auer, M., Zemla, M., Bill, M., Conrad, M. E., 2012. Deep-sea bacteria enriched by oil and dispersant from the Deepwater Horizon spill. Environmental microbiology. 14, 2405-2416.

Balakrishnan, G., Déniel, M., Nicolai, T., Chassenieux, C., Lagarde, F., 2019. Towards more realistic reference microplastics and nanoplastics: preparation of polyethylene micro/nanoparticles with a biosurfactant. Environmental Science: Nano. 6, 315-324.

Bini, S. A., 2018. Artificial intelligence, machine learning, deep learning, and cognitive computing: what do these terms mean and how will they impact health care? The Journal of arthroplasty. 33, 2358-2361.

Brakstad, O. G., Lewis, A., Beegle-Krause, C., 2018. A critical review of marine snow in the context of oil spills and oil spill dispersant treatment with focus on the Deepwater Horizon oil spill. Marine pollution bulletin. 135, 346-356.

Badem, H., Turkusagi, D., Caliskan, A., Çil, Z. A. 2019. Feature Selection Based on Artificial Bee Colony for Parkinson Disease Diagnosis. 2019 Medical Technologies Congress (TIPTEKNO), pp. 1-4.

Bang, D., Shim, H., 2021. MGGAN: Solving Mode Collapse Using Manifold-Guided Training, Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2347-2356.

Brandon, J., Goldstein, M., Ohman, M.D., 2016. Long-term aging and degradation of microplastic particles: comparing in situ oceanic and experimental weathering patterns. Marine pollution bulletin 110(1), 299-308.

Bayable, G., Amare, G., Alemu, G., Gashaw, T., 2021. Spatiotemporal variability and trends of rainfall and its association with Pacific Ocean Sea surface temperature in West Harerge Zone, Eastern Ethiopia. Environmental Systems Research. 10, 7.

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W. H., Ng, L. G., Ginhoux, F., Newell, E. W., 2019. Dimensionality reduction for visualizing single-cell data using UMAP. Nature Biotechnology. 37, 38-44.

Becker, M., Lippel, J., Stuhlsatz, A., Zielke, T., 2020. Robust dimensionality reduction for data visualization with deep neural networks. Graphical Models. 108, 101060.

Bhatnagar, S., Ghosal, D., Kolekar, M. H., Classification of fashion article images using convolutional neural networks. 2017 Fourth International Conference on Image Information Processing (ICIIP). IEEE, 2017, pp. 1-6.

Bzdok, D., Altman, N., Krzywinski, M., 2018. Points of significance: statistics versus machine learning. Nature Publishing Group.

Boucher, J. and Friot D. (2017). Primary Microplastics in the Oceans: A Global Evaluation of Sources. Gland, Switzerland: IUCN. 43pp.

Carroll, W. K., Huijzer, J., 2018. Who Owns Canada? s Fossil-Fuel Sector? Canadian Centre for Policy Alternatives.

Chang, S. E., Stone, J., Demes, K., Piscitelli, M., 2014. Consequences of oil spills: a review and approach for informing planning. Ecology and Society. 19.

Chen, R.-C., Dewi, C., Huang, S.-W., Caraka, R. E., 2020. Selecting critical features for data classification based on machine learning methods. Journal of Big Data. 7, 1-26.

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., Van Calster, B., 2019. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. Journal of Clinical Epidemiology. 110, 12-22.

Chen, G., Fu, Z., Yang, H., Wang, J., 2020. An overview of analytical methods for detecting microplastics in the atmosphere. TrAC Trends in Analytical Chemistry. 115981.

Chen, J., Zhang, W., Wan, Z., Li, S., Huang, T., Fei, Y., 2019. Oil spills from global tankers: Status review and future governance. Journal of cleaner production. 227, 20-32.

Cho, Y., Na, J.-G., Nho, N.-S., Kim, S., Kim, S., 2012. Application of saturates, aromatics, resins, and asphaltenes crude oil fractionation for detailed chemical characterization of heavy crude oils by Fourier transform ion cyclotron resonance mass spectrometry equipped with atmospheric pressure photoionization. Energy & Fuels. 26, 2558-2565.

Chu, X., Ilyas, I. F., Krishnan, S., Wang, J., Data cleaning: Overview and emerging challenges. Proceedings of the 2016 international conference on management of data, 2016, pp. 2201-2206.

Cincinelli, A., Scopetani, C., Chelazzi, D., Lombardini, E., Martellini, T., Katsoyiannis, A., Fossi, M. C., Corsolini, S., 2017. Microplastic in the surface waters of the Ross Sea (Antarctica): occurrence, distribution and characterization by FTIR. Chemosphere. 175, 391-400.

Clayton, J. R., Payne, J. R., Farlow, J. S., Sarwar, C., 2020. Oil spill dispersants: mechanisms of action and laboratory tests. CRC Press.

Chen, J.-C., Wang, Y.-M., 2020. Comparing Activation Functions in Modeling Shoreline Variation Using Multilayer Perceptron Neural Network. Water 12(5), 1281.

Chen, W., Hays, J., 2018. Sketchygan: Towards diverse and realistic sketch to image synthesis, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9416-9425.

Chen, Y., Chen, B., Song, X., Kang, Q., Ye, X., Zhang, B., 2021. A data-driven binary-classification approach for oil fingerprinting analysis. Environmental Research, 111454.

Corami, F., Rosso, B., Roman, M., Picone, M., Gambaro, A., Barbante, C., 2020. Evidence of small microplastics (< 100 μm) ingestion by Pacific oysters (Crassostrea gigas): A novel method of extraction, purification, and analysis using Micro-FTIR. Marine Pollution Bulletin. 160, 111606.

Cura, T., 2020. Use of support vector machines with a parallel local search algorithm for data classification and feature selection. Expert Systems with Applications. 145, 113133.

Czarnowski, I., Jędrzejowicz, P., Data reduction algorithm for machine learning and data mining. International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. Springer, 2008, pp. 276-285.

Dave, D., Ghaly, A. E., 2011. Remediation technologies for marine oil spills: a critical review and comparative analysis. American Journal of Environmental Sciences. 7, 423.

Dehghani, S., Moore, F., Akhbarizadeh, R., 2017. Microplastic pollution in deposited urban dust, Tehran metropolis, Iran. Environmental Science and Pollution Research. 24, 20360-20371.

Dewa, C. K., 2018. Suitable CNN weight initialization and activation function for Javanese vowels classification. Procedia computer science. 144, 124-132.

Dini, L., Mariano, S., Tacconi, S., Fidaleo, M., Rossi, M., 2021. Micro and Nanoplastics identification: Classic methods and innovative detection techniques. Frontiers in Toxicology. 3, 2.

Datta, A. R., Kang, Q., Chen, B., Ye, X., Chapter Four - Fate and Transport Modelling of Emerging Pollutants from Watersheds to Oceans: A Review. In: B. Chen, et al., Eds.), Advances in Marine Biology. Academic Press, 2018, pp. 97-128.

De Leonardis, G., Rosati, S., Balestra, G., Agostini, V., Panero, E., Gastaldi, L., Knaflitz, M., Human Activity Recognition by Wearable Sensors: Comparison of different classifiers for

real-time applications. 2018 IEEE International Symposium on Medical Measurements and Applications (MeMeA). IEEE, 2018, pp. 1-6.

Dewa, C.K., 2018. Suitable CNN weight initialization and activation function for Javanese vowels classification. Procedia computer science 144, 124-132.

Duan, J., Li, S.E., Guan, Y., Sun, Q., Cheng, B., 2020. Hierarchical reinforcement learning for self-driving decision-making without reliance on labelled driving data. IET Intelligent Transport Systems 14(5), 297-305.

Chicco, D., Jurman, G., 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC genomics 21(1), 1-13.

Dong, X. L., Rekatsinas, T., Data integration and machine learning: A natural synergy. Proceedings of the 2018 international conference on management of data, 2018, pp. 1645-1650.

Everaert, G., De Rijcke, M., Lonneville, B., Janssen, C., Backhaus, T., Mees, J., van Sebille, E., Koelmans, A., Catarino, A. I., Vandegehuchte, M. B., 2020. Risks of floating microplastic in the global ocean. Environmental Pollution. 267, 115499.

Firdaus, M., Trihadiningrum, Y., Lestari, P., 2020. Microplastic pollution in the sediment of Jagir Estuary, Surabaya City, Indonesia. Marine pollution bulletin. 150, 110790.

Fang, W., Zhang, F., Sheng, V.S., Ding, Y., 2018. A method for improving CNN-based image recognition using DCGAN. Computers, Materials and Continua 57(1), 167-178.

Gao, X., Pishdad-Bozorgi, P., 2020. A approach of developing machine learning models for facility life-cycle cost analysis. Building Research & Information. 48, 501-525.

Garreau, D., Luxburg, U., 2020. Explaining the explainer: A first theoretical analysis of LIME, International Conference on Artificial Intelligence and Statistics. PMLR, pp. 1287-1296.

Goldstein, J.I., Newbury, D.E., Michael, J.R., Ritchie, N.W., Scott, J.H.J., Joy, D.C., 2017. Scanning electron microscopy and X-ray microanalysis. Springer.

García, S., Luengo, J., Herrera, F., 2015. Data preprocessing in data mining. Springer.

Gaston, E., Woo, M., Steele, C., Sukumaran, S., Anderson, S., 2020. Microplastics differ between indoor and outdoor air masses: insights from multiple microscopy methodologies. Applied spectroscopy. 74, 1079-1098.

Golbayani, P., Florescu, I., Chatterjee, R., 2020. A comparative study of forecasting corporate credit ratings using neural networks, support vector machines, and decision trees. The North American Journal of Economics and Finance. 54, 101251.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Machine learning basics. Deep learning. 1, 98-164.

Gracia, A., Murawski, S. A., Vázquez-Bader, A. R., Impacts of deep oil spills on fish and fisheries. Deep oil spills. Springer, 2020, pp. 414-430.

Granata, F., Papirio, S., Esposito, G., Gargano, R., De Marinis, G., 2017. Machine learning algorithms for the forecasting of wastewater quality indicators. Water. 9, 105.

Grira, N., Crucianu, M., Boujemaa, N., 2004. Unsupervised and semi-supervised clustering: a brief survey. A review of machine learning techniques for processing multimedia content. 1, 9-16.

Guo, L.-Z., Zhou, Z., Li, Y.-F., Record: Resource constrained semi-supervised learning under distribution shift. Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 1636-1644.

Güven, O., Gökdağ, K., Jovanović, B., Kıdeyş, A. E., 2017. Microplastic litter composition of the Turkish territorial waters of the Mediterranean Sea, and its occurrence in the gastrointestinal tract of fish. Environmental pollution. 223, 286-294.

Harrison, J. P., Ojeda, J. J., Romero-González, M. E., 2012. The applicability of reflectance micro-Fourier-transform infrared spectroscopy for the detection of synthetic microplastics in marine sediments. Science of the Total Environment. 416, 455-463.

He, S., Wang, C., Li, Y., Yu, H., Han, B., 2016. Evaluation of diagnostic ratios of medium and serious weathered oils from five different oil sources. Acta Oceanologica Sinica. 35, 1-8.

Huang, J., Li, Y.-F., Xie, M., 2015. An empirical analysis of data preprocessing for machine learning-based software cost estimation. Information and software Technology. 67, 108-127.

Huang, Y., Gao, Z., Zhang, H., 2021. Comparison of common machine learning algorithms trained with multi-zone models for identifying the location and strength of indoor pollutant sources. Indoor and Built Environment. 30, 1142-1158.

Ilyas, I. F., Chu, X., 2019. Data cleaning. Morgan & Claypool.

Inoue, J., Yamagata, Y., Chen, Y., Poskitt, C. M., Sun, J., Anomaly detection for a water treatment system using unsupervised machine learning. 2017 IEEE International Conference on Data Mining Workshops (ICDMW). IEEE, 2017, pp. 1058-1065.

Jahnke, A., Arp, H. P. H., Escher, B. I., Gewert, B., Gorokhova, E., Kühnel, D., Ogonowski, M., Potthoff, A., Rummel, C., Schmitt-Jansen, M., 2017. Reducing uncertainty and confronting ignorance about the possible impacts of weathering plastic in the marine environment. Environmental Science & Technology Letters. 4, 85-90.

Jiang, T., Gradus, J. L., Rosellini, A. J., 2020. Supervised machine learning: a brief primer. Behavior Therapy. 51, 675-687.

Jiang, Y., Cukic, B., Menzies, T., Can data transformation help in the detection of fault-prone modules? , Proceedings of the 2008 workshop on Defects in large software systems, 2008, pp. 16-20.

Jöckel, L., Kläs, M., Martínez-Fernández, S., Safe traffic sign recognition through data augmentation for autonomous vehicles software. 2019 IEEE 19th International Conference on Software Quality, Reliability and Security Companion (QRS-C). IEEE, 2019, pp. 540-541.

Joo, C., Shim, W. J., Kim, G. B., Ha, S. Y., Kim, M., An, J. G., Kim, E., Kim, B., Jung, S. W., Kim, Y.-O., 2013. Mesocosm study on weathering characteristics of Iranian Heavy crude oil with and without dispersants. Journal of hazardous materials. 248, 37-46.

Jordan, M. I., Mitchell, T. M., 2015. Machine learning: Trends, perspectives, and prospects. Science. 349, 255-260.

Kara, K., Eguro, K., Zhang, C., Alonso, G., 2018. ColumnML: Column-store machine learning with on-the-fly data transformation. Proceedings of the VLDB Endowment. 12, 348-361.

Kim, W., Kanezaki, A., Tanaka, M., 2020. Unsupervised learning of image segmentation based on differentiable feature clustering. IEEE Transactions on Image Processing. 29, 8055-8068.

Kong, D., Zhu, J., Duan, C., Lu, L., Chen, D., 2020. Bayesian linear regression for surface roughness prediction. Mechanical Systems and Signal Processing. 142, 106770.

Kotsiantis, S. B., Kanellopoulos, D., Pintelas, P. E., 2006. Data preprocessing for supervised leaning. International journal of computer science. 1, 111-117.

Krishnan, S., Franklin, M. J., Goldberg, K., Wang, J., Wu, E., Activeclean: An interactive data cleaning approach for modern machine learning. Proceedings of the 2016 International Conference on Management of Data, 2016, pp. 2117-2120.

Kang, Q., Song, X., Xin, X., Chen, B., Chen, Y., Ye, X., Zhang, B., 2021. Machine Learning-Aided Causal Inference Approach for Environmental Data Analysis: A COVID-19 Case Study. Environmental Science & Technology.

Kuo, C., Yang, T., Kao, H., Wang, C., Lan, W., Tseng, H., 2018. Improvement of Envisat Altimetric Measurements in Taiwan Coastal Oceans by a Developed Waveform Retracking System. Journal of Environmental Informatics 31(1).

Kumar, V., Kalitin, D., Tiwari, P., Unsupervised learning dimensionality reduction algorithm PCA for face recognition. 2017 international conference on computing, communication and automation (ICCCA). IEEE, 2017, pp. 32-37.

Li, J., Qu, X., Su, L., Zhang, W., Yang, D., Kolandhasamy, P., Li, D., Shi, H., 2016. Microplastics in mussels along the coastal waters of China. Environmental pollution. 214, 177-184.

Li, Y., Wu, F.-X., Ngom, A., 2018. A review on machine learning principles for multi-view biological data integration. Briefings in bioinformatics. 19, 325-340.

Lim, S. K., Loo, Y., Tran, N.-T., Cheung, N.-M., Roig, G., Elovici, Y., Doping: Generative data augmentation for unsupervised anomaly detection with gan. 2018 IEEE International Conference on Data Mining (ICDM). IEEE, 2018, pp. 1122-1127.

Liu, R., Wei, S., Zhao, Y., Yang, Y., 2018. Indexing of the CNN features for the large scale image search. Multimedia Tools and Applications 77(24), 32107-32131.

Luo, G., Zhao, X., Tong, Y., Chen, Q., Zhu, Z., Lei, H., Lin, J., 2020. Geometry Sampling for 3D Face Generation via DCGAN, 2020 International Joint Conference on Neural Networks (IJCNN). IEEE, pp. 1-7.

Liu, P., Zhan, X., Wu, X., Li, J., Wang, H., Gao, S., 2020. Effect of weathering on environmental behavior of microplastics: Properties, sorption and potential risks. Chemosphere. 242, 125193.

Loh, W. Y., 2011. Classification and regression trees. Wiley interdisciplinary reviews: data mining and knowledge discovery. 1, 14-23.

Luo, Y., Lu, B.-L., EEG data augmentation for emotion recognition using a conditional Wasserstein GAN. 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2018, pp. 2535-2538.

Magnier, L., Mugge, R., Schoormans, J., 2019. Turning ocean garbage into products–Consumers' evaluations of products made of recycled ocean plastic. Journal of cleaner production. 215, 84-98.

Major, D., Zhang, Q., Wang, G., Wang, H., 2012. Oil-dispersant mixtures: understanding chemical composition and its relation to human toxicity. Toxicological & Environmental Chemistry. 94, 1832-1845.

Mansuy, L., Philp, R. P., Allen, J., 1997. Source identification of oil spills based on the isotopic composition of individual components in weathered oil samples. Environmental science & technology. 31, 3417-3425.

Magnier, L., Mugge, R., Schoormans, J., 2019. Turning ocean garbage into products–Consumers' evaluations of products made of recycled ocean plastic. Journal of cleaner production 215, 84-98.

Merlin, F., Zhu, Z., Yang, M., Chen, B., Lee, K., Boufadel, M.C., Isaacman, L., Zhang, B., 2021. Dispersants as marine oil spill treating agents: a review on mesoscale tests and field trials. Environmental Systems Research 10(1), 1-19.

Mescheder, L., Geiger, A., Nowozin, S., 2018. Which training methods for GANs do actually converge?, International conference on machine learning. PMLR, pp. 3481-3490.

Mikołajczyk, A., Grochowski, M., Data augmentation for improving deep learning in image classification problem. 2018 international interdisciplinary PhD workshop (IIPhDW). IEEE, 2018, pp. 117-122.

Mirnaghi, F. S., Pinchin, N. P., Yang, Z., Hollebone, B. P., Lambert, P., Brown, C. E., 2019. Monitoring of polycyclic aromatic hydrocarbon contamination at four oil spill sites using fluorescence spectroscopy coupled with parallel factor-principal component analysis. Environmental Science: Processes & Impacts. 21, 413-426.

Mulabagal, V., Yin, F., John, G., Hayworth, J., Clement, T., 2013. Chemical fingerprinting of petroleum biomarkers in Deepwater Horizon oil spill samples collected from Alabama shoreline. Marine Pollution Bulletin. 70, 147-154.

Nasteski, V., 2017. An overview of the supervised machine learning methods. Horizons. b. 4, 51-62.

Nguyen, N. B., Kim, M.-K., Le, Q. T., Ngo, D. N., Zoh, K.-D., Joo, S.-W., 2021. Spectroscopic analysis of microplastic contaminants in an urban wastewater treatment plant from Seoul, South Korea. Chemosphere. 263, 127812.

Ostle, C., Thompson, R. C., Broughton, D., Gregory, L., Wootton, M., Johns, D. G., 2019. The rise in ocean plastics evidenced from a 60-year time series. Nature communications. 10, 1-6.

Ougiaroglou, S., Diamantaras, K. I., Evangelidis, G., 2018. Exploring the effect of data reduction on Neural Network and Support Vector Machine classification. Neurocomputing. 280, 101-110.

Ozbayoglu, A.M., Gudelek, M.U., Sezer, O.B., 2020. Deep learning for financial applications: A survey. Applied Soft Computing 93, 106384.

Palinkas, L. A., 2012. A conceptual approach for understanding the mental health impacts of oil spills: lessons from the Exxon Valdez oil spill. Psychiatry: Interpersonal & Biological Processes. 75, 203-222.

Porter, A., Lyons, B. P., Galloway, T. S., Lewis, C., 2018. Role of marine snows in microplastic fate and bioavailability. Environmental science & technology. 52, 7111-7119.

Qassim, H., Verma, A., Feinzimer, D., 2018. Compressed residual-VGG16 CNN model for big data places image recognition, 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC). IEEE, pp. 169-175.

Qian, S., Liu, H., Liu, C., Wu, S., San Wong, H., 2018. Adaptive activation functions in convolutional neural networks. Neurocomputing 272, 204-212.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., Dormann, C. F., 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography. 40, 913-929.

Roch, S., Brinker, A., 2017. Rapid and efficient method for the detection of microplastic in the gastrointestinal tract of fishes. Environmental Science & Technology. 51, 4522-4530.

Ronkay, F., Molnar, B., Gere, D., Czigany, T., 2021. Plastic waste from marine environment: Demonstration of possible routes for recycling by different manufacturing technologies. Waste Management. 119, 101-110.

Sawyer, D., Stiebert, S., 2010. Fossil Fuels-At What Cost? Government Support for Upstream Oil Activities in Three Canadian Provinces: Alberta, Saskatchewan and Newfoundland & Labrador. Government Support for Upstream Oil Activities in Three Canadian Provinces: Alberta, Saskatchewan and Newfoundland & Labrador (November 2, 2010).

Scircle, A., Cizdziel, J. V., 2019. Detecting and quantifying microplastics in bottled water using fluorescence microscopy: A new experiment for instrumental analysis and environmental chemistry courses. Journal of Chemical Education. 97, 234-238.

Sheykhmousa, M., Mahdianpari, M., Ghanbari, H., Mohammadimanesh, F., Ghamisi, P., Homayouni, S., 2020. Support vector machine vs. 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.

Shim, W. J., Hong, S. H., Eo, S. E., 2017. Identification methods in microplastic analysis: a review. Analytical methods. 9, 1384-1391.

Sai, T.A., Lee, H.-h., 2018. Weight initialization on neural network for neuro pid controller-case study, 2018 International Conference on Information and Communication Technology Robotics (ICT-ROBOT). IEEE, pp. 1-4.

Sajjad, M., Zahir, S., Ullah, A., Akhtar, Z., Muhammad, K., 2020. Human behavior understanding in big multimedia data using CNN based facial expression recognition. Mobile networks and applications 25(4), 1611-1621.

Santurkar, S., Tsipras, D., Ilyas, A., Mądry, A., 2018. How does batch normalization help optimization?, Proceedings of the 32nd international conference on neural information processing systems. pp. 2488-2498.

Shan, J., Wang, J., Zhan, J., Liu, L., Wu, F., Wang, X., 2020. Sorption behaviors of crude oil on polyethylene microplastics in seawater and digestive tract under simulated real-world conditions. Chemosphere 257, 127225.

Shen, T., Zhou, T., Long, G., Jiang, J., Pan, S., Zhang, C., 2018. Disan: Directional self-attention network for rnn/cnn-free language understanding, Proceedings of the AAAI conference on artificial intelligence.

Shorten, C., Khoshgoftaar, T.M., 2019. A survey on image data augmentation for deep learning. Journal of Big Data 6(1), 1-48.

Sinaga, K. P., Yang, M.-S., 2020. Unsupervised K-means clustering algorithm. IEEE Access. 8, 80716-80727.

Song, X., Dispersed oil fingerprinting in marine environments. Memorial University of Newfoundland, 2019.

Su, Q., Liao, X., Li, C., Gan, Z., Carin, L., 2017. Unsupervised learning with truncated gaussian graphical models, Proceedings of the AAAI Conference on Artificial Intelligence.

Song, X., Lye, L. M., Chen, B., Zhang, B., 2019. Differentiation of weathered chemically dispersed oil from weathered crude oil. Environmental monitoring and assessment. 191, 1-10.

Song, X., Zhang, B., Chen, B., Cai, Q., 2016. Use of sesquiterpanes, steranes, and terpanes for forensic fingerprinting of chemically dispersed oil. Water, Air, & Soil Pollution. 227, 1-15.

Song, X., Zhang, B., Chen, B., Lye, L., Li, X., 2018. Aliphatic and aromatic biomarkers for fingerprinting of weathered chemically dispersed oil. Environmental Science and Pollution Research. 25, 15702-15714.

Stout, S., Wang, Z., 2016. Standard handbook oil spill environmental forensics: fingerprinting and source identification. Academic press.

Tollefson, J., Frickel, S., Restrepo, M. I., 2021. Feature extraction and machine learning techniques for identifying historic urban environmental hazards: New methods to locate lost fossil fuel infrastructure in US cities. Plos one. 16, e0255507.

Theckedath, D., Sedamkar, R., 2020. Detecting Affect States Using VGG16, ResNet50 and SE-ResNet50 Networks. SN Computer Science 1(2), 1-7.

Tu, Z., Xie, W., Zhang, D., Poppe, R., Veltkamp, R.C., Li, B., Yuan, J., 2019. A survey of variational and CNN-based optical flow techniques. Signal Processing: Image Communication 72, 9-24.

Vallecorsa, S., Carminati, F., Khattak, G., 3D convolutional GAN for fast simulation. EPJ Web of Conferences, Vol. 214. EDP Sciences, 2019, pp. 02010.

Veerasingam, S., Ranjani, M., Venkatachalapathy, R., Bagaev, A., Mukhanov, V., Litvinyuk, D., Mugilarasan, M., Gurumoorthi, K., Guganathan, L., Aboobacker, V., 2021. Contributions of Fourier transform infrared spectroscopy in microplastic pollution research: A review. Critical Reviews in Environmental Science and Technology. 51, 2681-2743.

Visani, G., Bagli, E., Chesani, F., Poluzzi, A., Capuzzo, D., 2020. Statistical stability indices for LIME: obtaining reliable explanations for machine learning models. Journal of the Operational Research Society, 1-11.

Volokitin, A., Konukoglu, E., Van Gool, L., 2020. Decomposing image generation into layout prediction and conditional synthesis, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 372-373.

Waheed, A., Goyal, M., Gupta, D., Khanna, A., Al-Turjman, F., Pinheiro, P. R., 2020. Covidgan: data augmentation using auxiliary classifier gan for improved covid-19 detection. Ieee Access. 8, 91916-91923.

Wang, Z., Stout, S., 2010. Oil spill environmental forensics: fingerprinting and source identification. Elsevier.

Worthington, M.J., Shearer, C.J., Esdaile, L.J., Campbell, J.A., Gibson, C.T., Legg, S.K., Yin, Y., Lundquist, N.A., Gascooke, J.R., Albuquerque, I.S., 2018. Sustainable polysulfides for oil

spill remediation: repurposing industrial waste for environmental benefit. Advanced Sustainable Systems 2(6), 1800024.

Wu, H., Huang, Q., Wang, D., Gao, L., 2018. A CNN-SVM combined model for pattern recognition of knee motion using mechanomyography signals. Journal of Electromyography and Kinesiology 42, 136-142.

Wang, Z., Stout, S. A., Fingas, M., 2006. Forensic fingerprinting of biomarkers for oil spill characterization and source identification. Environmental Forensics. 7, 105-146.

Xin, X., Huang, G., An, C., Feng, R., 2019. Interactive toxicity of triclosan and nano-TiO2 to green alga Eremosphaera viridis in Lake Erie: A new perspective based on Fourier transform infrared spectromicroscopy and synchrotron-based X-ray fluorescence imaging. Environmental science & technology 53(16), 9884-9894.

Yang, M., Chen, B., Xin, X., Song, X., Liu, J., Dong, G., Lee, K., Zhang, B., 2021. Interactions between microplastics and oil dispersion in the marine environment. Journal of Hazardous Materials. 403, 123944.

Yang, Z., Yang, C., Wang, Z., Hollebone, B., Landriault, M., Brown, C. E., 2011. Oil fingerprinting analysis using commercial solid phase extraction (SPE) cartridge and gas chromatography-mass spectrometry (GC-MS). Analytical Methods. 3, 628-635.

Yang, M., Chen, B., Xin, X., Song, X., Liu, J., Dong, G., Lee, K., Zhang, B., 2021. Interactions between microplastics and oil dispersion in the marine environment. Journal of Hazardous Materials 403, 123944.

Ye, X., Chen, B., Lee, K., Storesund, R., Li, P., Kang, Q., Zhang, B., 2021. An emergency response system by dynamic simulation and enhanced particle swarm optimization and application for a marine oil spill accident. Journal of Cleaner Production 297, 126591.

Zhao, S., Danley, M., Ward, J. E., Li, D., Mincer, T. J., 2017. An approach for extraction, characterization and quantitation of microplastic in natural marine snow using Raman microscopy. Analytical methods. 9, 1470-1478.

Zhong, S., Zhang, K., Bagheri, M., Burken, J. G., Gu, A., Li, B., Ma, X., Marrone, B. L., Ren, Z. J., Schrier, J., 2021. Machine Learning: New Ideas and Tools in Environmental Science and Engineering. Environmental Science & Technology. 55, 12741-12754.

Zhong, Y. D., Leonard, N., 2020. Unsupervised learning of lagrangian dynamics from images for prediction and control. Advances in Neural Information Processing Systems. 33.

Zoph, B., Cubuk, E. D., Ghiasi, G., Lin, T.-Y., Shlens, J., Le, Q. V., Learning data augmentation strategies for object detection. European Conference on Computer Vision. Springer, 2020, pp. 566-583.

Zhang, Z., Beck, M.W., Winkler, D.A., Huang, B., Sibanda, W., Goyal, H., 2018. Opening the black box of neural networks: methods for interpreting neural network models in clinical applications. Annals of translational medicine 6(11).

Zhang, Z., Yu, J., 2019. STDGAN: ResBlock Based Generative Adversarial Nets Using Spectral Normalization and Two Different Discriminators, Proceedings of the 27th ACM International Conference on Multimedia. pp. 674-682.

Zhu, Q., Du, B., Turkbey, B., Choyke, P.L., Yan, P., 2017. Deeply-supervised CNN for prostate segmentation, 2017 international joint conference on neural networks (IJCNN). IEEE, pp. 178-184.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., He, Q., 2020. A comprehensive survey on transfer learning. Proceedings of the IEEE 109(1), 43-76.

APPENDIX

Table A1 Characteristics of biomarkers in the thesis (Song et al., 2019)

Peak	Compounds	Formula	Target ions
Diamantanes			
1	diamantane	C14H20	188
2	4-methyldiamantane	C15H22	187
3	4,9-dimethyldiamantane	C16H24	201
4	1-methyldiamantane	C15H22	187
5	1,4-and 2,4-dimethyldiamantane	C16H24	201
6	4,8-dimethyldiamantane	C16H24	201
7	trimethyldiamantane	C17H26	215
8	3-methyldiamantane	C15H22	187
9	3,4-dimethyldiamantane	C16H24	201
Steranes			
DIA27S (1)	C27 20S-13β(H),17α(H)-diasterane	C27H46	217, 218
DIA27R (2)	C27 20R-13β(H),17α(H)-diasterane	C27H46	217, 218
C27S (7)	C27 20S-5α(H),14α(H),17α(H)-cholestane	C27H48	217, 218
C27αββR (8)	C27 20R-5α(H),14β(H),17β(H)-cholestane	C27H48	217, 218
C27αββS (10)	C27 20S-5α(H),14β(H),17β(H)-cholestane	C27H48	217, 218

Table A1 Characteristics of biomarkers in the thesis (Song et al., 2019)

Peak	Compounds	Formula	Target ions
C27R (11)	C27 20R-5α(H),14α(H),17α(H)-cholestane	C27H48	217, 218
C28S(13)	C28 20S-5α(H),14α(H),17α(H)-ergostane	C28H50	217, 218
C28αββR(14)	C28 20R-5α(H),14β(H),17β(H)-ergostane	C28H50	217, 218
C28αββS(15)	C28 20S-5α(H),14β(H),17β(H)-ergostane	C28H50	217, 218
C28R(16)	C28 20R-5α(H),14α(H),17α(H)-ergostane	C28H50	217, 218
C29S	C29 20S-5α(H),14α(H),17α(H)-stigmastane	C29H52	217, 218
C29αββR	C29 20R-5α(H),14β(H),17β(H)-stigmastane	C29H52	217, 218
C29αββS	C29 20S-5α(H),14β(H),17β(H)-stigmastane	C29H52	217, 218
C29R	C29 20R-5α(H),14α(H),17α(H)-stigmastane	C29H52	217, 218

Continued Table A1 Characteristics of biomarkers in the thesis (Song et al., 2019)

Peak	Compounds	Formula	Target ions
Terpanes			
C23	C23 tricyclic terpane	C23H42	191
C24	C24 tricyclic terpane	C24H44	191
C25	C25 tricyclic terpane (a)	C25H46	191
C26	C26 (S + R) tricyclic terpanes	C24H42 + C26H48	191
TR28a	C28 tricyclic terpane (a)	C28H52	191
TR28b	C28 tricyclic terpane (b)	C28H52	191
TR29a	C29 tricyclic terpane (a)	C29H54	191
TR29b	C29 tricyclic terpane (b)	C29H54	191
Ts	18α(H),21β(H)-22,29,30-trisnorhopane	C27H46	191
Tm	17α(H),21β(H)-22,29,30-trisnorhopane	C27H46	191
H29 (C29)	17α(H),21β(H)-30-norhopane	C29H50	191
C29TS	18α(H),21β(H)-30-norneohopane	C29H50	191
M29	17α(H),21β(H)-30-norhopane	C29H50	191
H30 (C30)	17α(H),21β(H)-hopane	C30H52	191
H31S (C31S)	22S-17α(H),21β(H)-30-homohopane	C31H54	191
H31R (C31R)	22R-17α(H),21β(H)-30-homohopane	C31H54	191

Continued Table A1 Characteristics of biomarkers in the thesis (Song et al., 2019)

Peak	Compounds	Formula	Target ions
H32S (C32S)	22S-17α(H),21β(H)-30,31-bishomohopane	C32H56	191
H32R (C32R)	22R-17α(H),21β(H)-30,31-bishomohopane	C32H56	191
H33S (C33S)	22S-17α(H),21β(H)-30,31,32-trishomohopane	C33H58	191
H33R (C33R)	22R-17α(H),21β(H)-30,31,32-trishomohopane	C33H58	191
H34S (C34S)	22S-17α(H),21β(H)-30,31,32,33-tetrakishomohopane	C34H60	191
H34R (C34R)	22R-17α(H),21β(H)-30,31,32,33-tetrakishomohopane	C34H60	191
H35S (C35S)	22S-17α(H),21β(H)-30,31,32,33,34-pentakishomohopane	C35H62	191
H35R (C35R)	22R-17α(H),21β(H)-30,31,32,33,34-pentakishomohopane	C35H62	191
TA-steranes			
1	C20 triaromatic-sterane	C10H16	231
2	C21 triaromatic-sterane	C11H18	231
3a	C22 triaromatic steroids (a)	C12H20	231

Continued Table A1 Characteristics of biomarkers in the thesis (Song et al., 2019)

Peak	Compounds	Formula	Target ions
3b	C22 triaromatic steroids (b)	C13H22	231
4	C26 triaromatic-chloestane (20S)	C14H24	231
5	C26 triaromatic-chloestane(20R) + C27triaromatic-ergostane(20S)	C11H18	231
6	C28 triaromatic-stigmastane (20S)	C12H20	231
7	C27 triaromatic-ergostane (20R)	C12H20	231
8	C28 triaromatic-stigmastane (20R)	C13H22	231
MA-steranes			
1	C21 5ß monoaromatic steroid	C21H30	253
2	C21 5a monoaromatic steroid	C21H30	253
3a	C23 monoaromatic steroid (20S)	C22H32	253
3b	C23 monoaromatic steroid (20R)	C22H32	253
4	C27 monoaromatic 5ß(H)-cholestane (20S)	C27H42	253
5	C27 monoaromatic diacholestane (20S)	C27H42	253

Continued Table A1 Characteristics of biomarkers in the thesis (Song et al., 2019)

Peak	Compounds	Formula	Target ions
7	(C27 monoaromatic 5α (H)-cholestane (20S)) + C28 monoaromatic 5ß(H)-ergostane(20S) + diaergostane (20S)	C27H42+ C28H44	253
8	C27 monoaromatic 5α (H)-cholestane (20R)	C27H42	253
9	C28 monoaromatic 5α (H)-ergostane (20S)	C28H44	253
10	C28 monoaromatic 5ß(H)-ergostane (20R) +diaergostane (20R)	C28H44	253
11	C29 monoaromatic 5α (H)-stigmastane (20S)	C29H46	253
12	C28 monoaromatic 5α (H)-ergostane (20R)	C28H44	253, 193

Table A2 Outputs of different ML models in all biomarkers

Biomarkers	Datasets	Models	Best model score	Best parameters
Terpanes	X1 (original dataset)	LRC	0.827	C:20, solver: liblinear
		RFC	0.864	n-estimators: 100
		DTC	0.845	Criterion: gini, splitter: best
		SVC	0.827	kernel: linear
		KNN	0.800	n_neighbors: 5
		EVC	0.870	Vote: soft
	X (all PCs)	LRC	0.827	C:20, solver: liblinear
		RFC	0.863	n-estimators: 100
		DTC	0.890	Criterion: gini, splitter: best
		SVC	0.842	kernel: linear
		KNN	0.727	n_neighbors: 10
		EVC	0.864	Vote: soft
	X3 (top 3 PCs)	LRC	0.718	C:1, solver: liblinear
		RFC	0.836	n-estimators: 100
		DTC	0.836	Criterion: entropy, splitter: random
		SVC	0.727	kernel: linear
		KNN	0.780	n_neighbors: 5
		EVC	0.855	Vote: soft
TA-steranes	X1 (original dataset)	LRC	0.825	C:30, solver: liblinear
		RFC	0.830	n-estimators: 100
		DTC	0.854	Criterion: gini, splitter: random
		SVC	0.783	kernel: linear
		KNN	0.878	n_neighbors: 5
		EVC	0.880	Vote: soft
	X (all PCs)	LRC	0.825	C:30, solver: liblinear
		RFC	0.875	n-estimators: 100
		DTC	0.841	Criterion: entropy, splitter: random
		SVC	0.783	kernel: linear
		KNN	0.683	n_neighbors: 7

		EVC	0.900	Vote: soft
	X3 (top 3 PCs)	LRC	0.817	C:1, solver: liblinear
		RFC	0.867	n-estimators: 50
		DTC	0.833	Criterion: gini, splitter: random
		SVC	0.842	kernel: rbf
		KNN	0.85	n_neighbors: 5
		EVC	0.858	Vote: soft
Steranes	X1 (original dataset)	LRC	0.727	C:1, solver: liblinear
		RFC	0.818	n-estimators: 100
		DTC	0.781	Criterion: gini, splitter: best
		SVC	0.736	kernel: linear
		KNN	0.780	n_neighbors: 5
		EVC	0.800	Vote: soft
	X (all PCs)	LRC	0.727	C:1, solver: liblinear
		RFC	0.845	n-estimators: 100
		DTC	0.800	Criterion: entropy, splitter: best
		SVC	0.736	kernel: rbf
		KNN	0.773	n_neighbors: 7
		EVC	0.800	Vote: soft
	X3 (top 3 PCs)	LRC	0.627	C:1, solver: liblinear
		RFC	0.709	n-estimators: 100
		DTC	0.782	Criterion: gini, splitter: best
		SVC	0.682	kernel: rbf
		KNN	0.718	n_neighbors: 5
		EVC	0.709	Vote: soft
MA-steranes	X1 (original dataset)	LRC	0.856	C:30, solver: liblinear
		RFC	0.836	n-estimators: 100
		DTC	0.816	Criterion: gini, splitter: random
		SVC	0.864	kernel: linear
		KNN	0.810	n_neighbors: 7
		EVC	0.856	Vote: soft
	X (all PCs)	LRC	0.704	C:1, solver: liblinear
		RFC	0.848	n-estimators: 100
		DTC	0.792	Criterion: gini, splitter: random

		SVC	0.864	kernel: rbf
		KNN	0.840	n_neighbors: 5
		EVC	0.832	Vote: soft
	X3 (top 3 PCs)	LRC	0.712	C:1, solver: liblinear
		RFC	0.824	n-estimators: 150
		DTC	0.800	Criterion: gini, splitter: random
		SVC	0.808	kernel: rbf
		KNN	0.834	n_neighbors: 5
		EVC	0.852	Vote: soft
Diamantanes	X1 (original dataset)	LRC	0.743	C:1, solver: liblinear
		RFC	0.829	n-estimators: 150
		DTC	0.800	Criterion: gini, splitter: random
		SVC	0.629	kernel: linear
		KNN	0.835	n_neighbors: 5
		EVC	0.857	Vote: soft
	X (all PCs)	LRC	0.686	C:1, solver: liblinear
		RFC	0.900	n-estimators: 30
		DTC	0.743	Criterion: gini, splitter: best
		SVC	0.771	kernel: rbf
		KNN	0.729	n_neighbors: 5
		EVC	0.686	Vote: soft
	X3 (top 3 PCs)	LRC	0.457	C:1, solver: liblinear
		RFC	0.771	n-estimators: 50
		DTC	0.829	Criterion: gini, splitter: random
		SVC	0.714	kernel: rbf
		KNN	0.757	n_neighbors: 7
		EVC	0.771	Vote: soft