Comparative Analysis between ResNet Models on Marine Oil spill detection

MSc Research Project
Data Analytics

Abin Joseph
Student ID: 21216312

School of Computing
National College of Ireland

Supervisor: Abdul Shahid

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Abin Joseph
Student ID:	21216312
Programme:	Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Abdul Shahid
Submission Due Date:	31/01/2024
Project Title:	Comparative Analysis between ResNet models on marine oil spill detection.
Word Count:	6139
Page Count:	18

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Abin Joseph
Date:	31/01/2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Comparative Analysis between ResNet models on marine oil spill detection

Abin Joseph
X21216312

Abstract

Marine oil spills remain a significant threat to ecosystems and maritime safety, necessitating advanced detection methodologies. Despite a decrease in the frequency of spills, challenges exist in timely and accurate detection. This research evaluates state of the art deep learning models, including ResNet50, ResNet50V2, and ResNet101, with the previously used VGG19 model for SAR image classification in oil spill detection. The study addresses misclassification issues, focusing on the effectiveness of these models in classifying SAR images with oil-like and non-oil-like features. The evaluation utilizes metrics such as the classification report, confusion matrix, and ROC curve. Results shows that ResNet50 outperforms other models, achieving a weighted F-score of 0.95 and a ROC curve area of 0.99. The research contributes valuable insights to environmental monitoring, emphasizing the potential replacement of VGG19 with ResNet50 for improved oil spill detection.

Keywords—Marine oil spills, SAR imagery, Deep learning models, ResNet50, Classification evaluation, Environmental monitoring

1 Introduction

Marine oil spills pose a serious threat to ecosystems and maritime safety (Zhang et al.,2019), demanding advanced methodologies for timely and accurate detection to effectively mitigate the impacts. As per the statistical reports the medium (between 7 to 700 tonnes) and large (greater than 700 tonnes) scale oil spills showed a considerable reduction in the previous decade from 71.7% (2000 to 2009) to 43.75% (2010 to 2019) which indicates the advancements in safety protocols and new types of ships (Chen et al., 2019). Although the number of oil spills has gone down in recent years, it is still difficult to detect, prevent and clean up the spills.

Numerous approaches have been explored in the scientific community to address the complexities of oil spill detection, ranging from manual methods to empirical and machine learning-based strategies. Manual detection, while thorough, faces limitations in terms of generalizability, particularly across diverse geological and environmental conditions. Empirical approaches and machine learning methods strives to speed up the detection process, but challenges continue in dataset quality and environmental factors.

This research contributes to the scientific literature by focusing on the evaluation of state of the art deep learning models for image classification like ResNet50, ResNet50V2, and ResNet101 in comparison with the previously employed VGG19 model (Blondeau-Patissier et al.,2023). The primary research question centers on the effectiveness of these models in classifying SAR images containing oil-like and non-oil-like features. Through a comprehensive examination of these models, the study aims to address existing misclassification issues and enhance the reliability of oil spill detection in SAR imagery. Figure1 shows the randomly chosen sample images from class 0 and class 1.

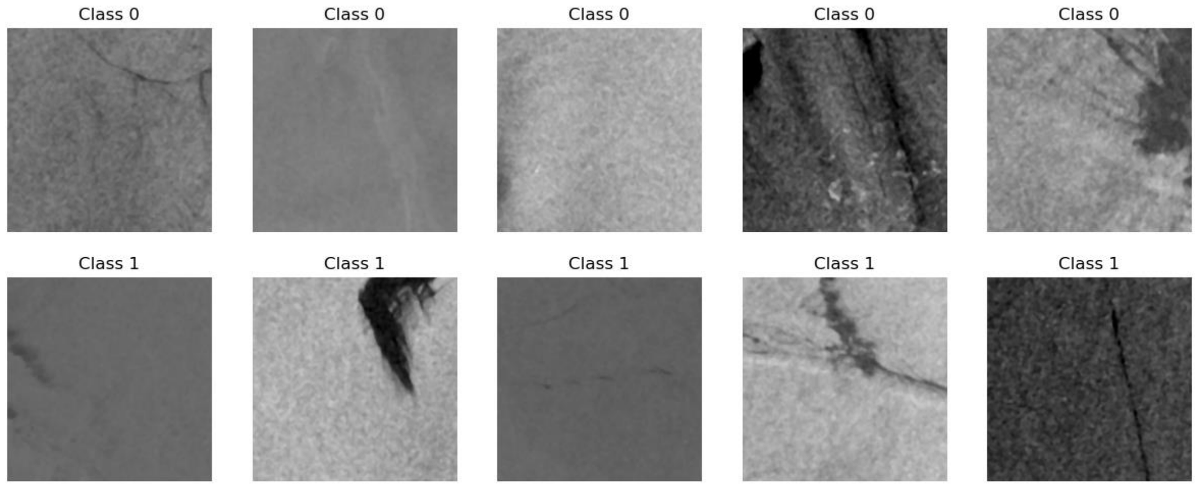

Figure 1. Samples from CSIRO Sentinel-1 SAR image dataset containing oil like (class 1) and non-oil like (class 0) features.

The report follows the given structure: Section 1.1 contains the research question this research trying to address. Section 2 provides an overview of related works, elucidating on manual, empirical, and machine learning-based approaches employed in oil spill detection. Section 3 outlines the research methodology, detailing the CRISP-DM approach, business understanding, data understanding, model building, and evaluation. Section 4 explains the implementation details, followed by Section 5, which evaluates the trained models. The discussion in Section 6 discussions interprets the results, and Section 7 concludes the study, presenting avenues for future research. This research not only contributes valuable insights to the field of environmental monitoring but also offers a detailed evaluation of deep learning models for SAR image classification in oil spill detection.

1.1 Research Question

How effectively can SAR images containing oil-like features and non-oil-like features be classified using ResNet50, ResNet50V2, and ResNet101 models compared to the proposed VGG19 model?

2 Related Works

Detecting oil slicks is a crucial aspect of monitoring the environment and ensuring maritime safety. The detection and addressing of undisclosed and unlawful oil discharges carry substantial consequences for both the marine ecosystem and the contemporary global environment. There are many approaches that has been tried for the effective classification of oil and non-oil like features from SAR images.

2.1 Manual detection of oil spills.

A multi-faceted approach, combining long-term and short-term monitoring with SAR imagery, geological analysis, correlation with wind speeds, and residence time estimation to comprehensively understand seepage dynamics in the Lower Congo Basin (Jatiault et al., 2017). This approach helps in identifying recurrent patterns and trends but lacks generalisability in other regions with different geological and environmental conditions.

Spatial density thresholding is used for the detection of dark spots in SAR imagery for oil-spill monitoring. This innovation contributes to the field by providing an alternative to more commonly used intensity-based methods (Shu et al.,2010). Even though this approach is fast and effective it does not perform well enough when the dark spots are not-well-defined, linear or are in a heterogeneous background. These manual detection methods show good results in the chosen area of study difficult to handle when large amounts of data are available.

2.2 Empirical based approaches to detect oil spills.

This approach came in being to reduce the time required for visual and manual detection of oil spills in SAR image. Basically, uses a rule-based method to differentiate between oil like features and non-oil like features. An automatic seepage location estimation method using SAR images was proposed which incorporates contextual wind information. It reduces the processing time enabling the detections on a large-scale data. The effectiveness of the method deeply relies on the availability and quality of SAR data. Cloud cover or data artifacts could affect the accuracy of the results (Suresh et al., 2017).

Using well defined feature extraction can be helpful to differentiate oil spills and other objects on water. Extracting the suitable and most optimum features like Geometric, Statistical, textual, contextual and polarimetric to classify oil like and non-oil like features is based on the experience of the researchers (Al-Ruzouq et al., 2020). Even though this constitutes for better classification it lacks generalisability since the suitable features can be varied according to the location of the spill.

2.3 ML and DL based approaches to detect oil spills.

With the usage of mRMR_SVM(minimum redundancy and maximum relevance) supervised algorithm to identify oil related features from SAR images where mRMR algorithm is used for feature selection which is helpful to reduce vector dimension. And SVM with RBF kernel is applied for classification (Zhou et al.,2018). Even though it is fast, adaptable and utilises dimensionality reduction technique the model can stumble upon real world scenarios.

To develop a monitoring system for the automatic detection of oil spill events caused by ships (bilge dumping) in African Oceans, using synthetic aperture radar (SAR) imagery (Mdakane and Kleynhans, 2020). The study focuses on discriminating oil spills from natural phenomena known as oil spill look-alikes, which can also dampen radar energy return and appear as linearly shaped dark regions in SAR images. They used multiple feature selection methods to determine critical features and to rank them. Then the selected features are used for classification using GBT classifier (Gradient Boosted Tree classifier). Feature engineering is a crucial aspect of the proposed method. The need for manual or algorithmic feature selection can be time-consuming, and there's a risk of not capturing all relevant information in the data.

Traditional NN and DL techniques works way beyond manual and empirical approaches while addressing environmental remote sensing (Yuan et al., 2020). In the proposed review it covered various aspects including DL architectures in environmental domains like atmosphere, vegetation, oceanography, hydrology etc. Usage of transfer learning while working with limited samples and inclusion of geographical laws was suggested. The review provides insights on utilising DL methods falls short on explaining the challenges, temporal conditions, and generalisation.

Deep learning model specifically Faster RCNN shows promising approach for the marine oil spill detection. This study utilises large, labelled dataset collected from C-Band Sentinal-1A/B and RADARSAT-2 SAR images for training, testing and validation (Huang et al., 2022). It is designed to achieve fast and effective oil spill detection by overcoming limitations on algorithm complexity, imbalanced datasets, and uncertainties in feature selection by reducing detection time. Even though it has many strengths as pointed out the availability of quality training data as well as environmental factors can affect the performance of this method which includes discriminating the oil spills from look-alikes.

With the help of DCNNs, semantic segmentation was used for efficient oil spill detection in SAR images. Where the primary goal is to address the challenges in discriminating oil spills from look-alikes (Krestenitis et al., 2019). The semantic segmentation gives detailed and accurate identification, but it requires pixel wise annotations on the dataset which can be expensive to acquire. They have made the dataset to be publicly available to serve as a common benchmark allowing fair and standardized evaluation. Semantic segmentation can be used in multiclassification of remote sensing image segmentation (Zheng and Chen, 2021) effectively this study uses binary segmentation. Due to the less availability of the annotated dataset, binary classification using deep learning rather than segmentation seems to be more flexible and cost effective.

The base paper selected for this study employed a combination of deep learning and empirical approaches to create a semi-automated detection system which gave a promising result, as documented by Blondeau-Patissier (Blondeau-Patissier et al., 2023), to identify oil slicks. In the research they used VGG19 pretrained model along with empirical methods since VGG19 gave a better Fscore of 0.9. Even though the proposed method gives good performance still require a trained operator to define threshold values and parameters. As per the confusion matrix given in the base paper shows that the VGG19 model used is misclassifying the classes. Mostly the minority class which contain images with oil like features. 31% of the images with oil like features were classified as False Negatives.

Table 1 contains the strength and weaknesses various ML and DL based approaches to detect oil spills as discussed in section 2.3.

Reference	Approach	Strengths	Weaknesses
Zhou et al., 2018	mRMR_SVM algorithm	- Utilizes dimensionality reduction techniques.	- May stumble upon real-world scenarios due to lack of Generalization
Mdakane and Kleynhans, 2020	Feature selection and GBT classifier	- Uses multiple feature selection methods like ANOVA and RFE.	-Manual or algorithmic feature selection can be time-consuming. - Risks not capturing all relevant information in the data.
Yuan et al., 2020	Traditional NN and DL techniques	- Covers various aspects of DL architectures in environmental domains by suggesting transfer learning for limited samples.	- Falls short on explaining challenges, temporal conditions, and generalization issues.

Huang et al., 2022	Faster RCNN with large, labeled dataset	- Overcomes limitations in algorithm complexity and imbalanced datasets.	- Performance may be affected by the availability of quality training data - Challenges in discriminating oil spills from look-alikes.
Krestenitis et al., 2019	DCNNs for semantic segmentation	- Employs semantic segmentation for detailed and accurate identification.	- Requires pixel-wise annotations, which can be expensive and time consuming.
Zheng and Chen, 2021	Semantic segmentation for binary classification	- More flexible and cost-effective than pixel-wise segmentation.	- Limited availability of annotated datasets for segmentation. - Looks like an over kill for binary classification
Blondeau-Patissier et al., 2023	Deep learning and empirical approaches	- Semi-automated detection system with a promising result. - Uses a pretrained VGG19 model.	- Requires a trained operator for defining threshold values and parameters since - VGG19 model misclassifies OLF and NOLF which decrease the performance of this approach.
Mascarenhas and Agarwal, 2021	Comparative analysis between VGG models and ResNet models	- ResNet50 outperformed VGG19	- Only considers train and test accuracy as the evaluation metrics.

Table 1. Strength and weakness of various studies over the year

In a comparative analysis between VGG models and ResNet models on an Image classification problem ResNet50 outperformed VGG19 and VGG16 (Mascarenhas and Agarwal, 2021). Since the ResNet model outperformed VGG19, ResNet models should be explored further with optimum number of epochs, learning rates, optimizing function and suitable data augmentation techniques.

So, in this study, various ResNet models like ResNet50, ResNet50V2 and ResNet101 are used to evaluate the performance on the SAR image dataset which is made available from the previous study (Blondeau-Patissier et al.,2022), by incorporating the actual image size and preprocessing steps. Since the dataset is skewed toward the majority class, appropriate data augmentation will be applied to avoid class imbalance.

The proposed model reduces the misclassification of the images, it could replace the VGG19 model with suggested ResNet model to further enhance the semi-automated detection system, thereby improving the efficiency of oil detection by reducing human dependability.

3 Research Methodology

This research was conducted by following CRISP-DM approach which includes business understanding, data understanding, data preparation, modelling, and evaluation. CRISP-DM process life cycle is shown on Figure 2.

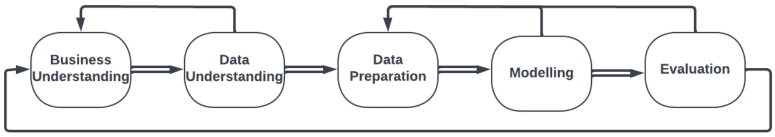

Figure 2. CRISP-DM approach used in this study.

From the Figure 2 it is evident that these phases are not strictly linear; they often involve iteration and revisiting previous stages based on insights gained during the project. The CRISP-DM approach provides a systematic and structured methodology for data mining projects, ensuring that each step contributes to the overall success of the analysis.

3.1 Business Understanding

The primary objective is to assess the efficiency of classifying SAR images containing oil-like features and non-oil-like features. The focus is on comparing the classification performance of ResNet50, ResNet50V2, and ResNet101 models against the currently proposed VGG19 model. This evaluation aims to inform decision-making processes related to the choice of the most effective deep learning model for accurately identifying oil-like features in SAR imagery, which is crucial for applications in environmental monitoring and oil spill detection.

3.2 Device Specification

Google Colaboratory was used for coding due to the availability of better computational resources such as A100 GPU with 50GB RAM, allocated System RAM with 83.5 GB and Disk space of 166.8 GB. The dataset was uploaded to the Google Drive in order make it accessible from Colab notebook. The code is developed using TensorFlow library as the backend and Keras as the high-level API to construct deep learning models due to their flexibility and ease of use.

3.3 Data Understanding

In this section the EDA will be applied on the dataset. The tools, frameworks and necessary libraries and system configuration used for model building and training will be discussed in detail.

3.3.1 Data collection

CSIRO Sentinel-1 SAR image dataset (Blondeau-Patissier et al., 2022) used for this study was downloaded from CSIRO data access portal which is useful for training and testing deep learning models for detecting oil like features in SAR images. It comes with Creative Commons Attribution-Share like 4.0 International License which allows user to share and adapt. The dataset contains images or image chips of size 400 * 400 pixels. Each image is labelled as 0 if it doesn’t contain oil like features and 1 if it contains oil like features.

3.3.2 Exploratory Data Analysis

EDA can give us many useful information that could be used in preprocessing, data augmentation, model training and evaluation.

Class Imbalance

The dataset is imbalanced and skewed to class 0 which is two by third of the entire dataset (66%) whereas class 0 corresponds to remaining one by third of the dataset (34%) as shown in Figure 3. That is, from the total of 5630 images 3725 belongs to class 0 and 1905 images belongs to class 1.

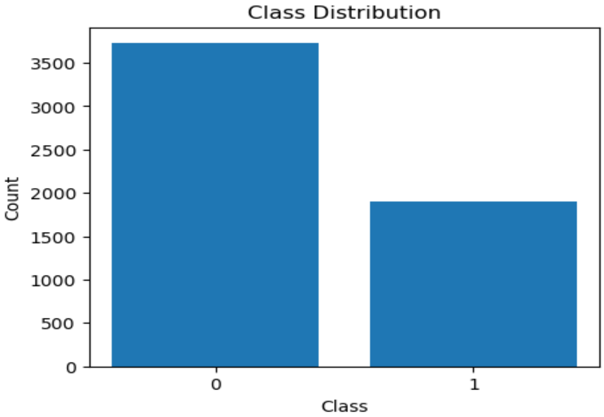

Figure 3. CSIRO Sentinel-1 SAR image dataset class distribution

Pixel Analysis

The mean pixel values of the oil (class 1) and non-oil slick (class 0) images shows that the images containing oil slicks tends to have less mean pixel value than the images with non-oil slicks as shown in Figure 4. On average, the pixels in oil slick images have lower intensity or brightness compared to pixels in non-oil slick images. Oil slicks appear darker in the images compared to the surrounding non-oil areas. This could be due to the physical properties of oil, which may absorb, or scatter light differently than the materials present in non-oil areas.

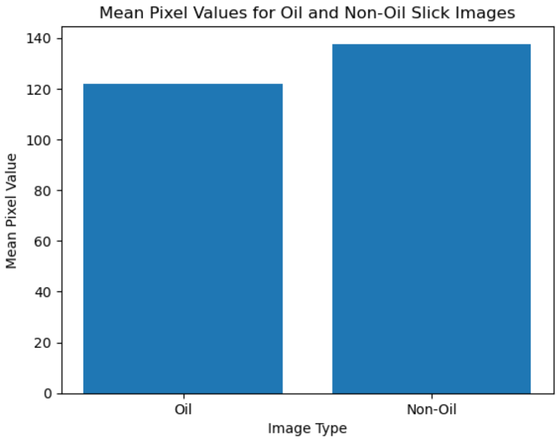

Figure 4. Mean pixel value of Oil (class 1) and non-oil slicks (class 0) in images.

3.4 Data Preparation

The analysed image data set will be further prepared for the training by effectively applying appropriate SAR image preprocessing and transformations, data augmentation and dividing the dataset in to Train, Test and validation directory.

3.4.1 SAR Image Preprocessing and transformations

During calibration the pixel values of the image is normalised in the range of [0,255] which is a common scale for images by using Min-max normalisation technique. This calibrated image is subjected to speckle reduction using a median blur filter. This filter replaces each pixel value with the median value of its neighbouring pixels which can effectively smoothen the image to reduce the impact of noise or speckle patterns. This speckle reduced image will be used for multilooking. Multilooking is a commonly used radar image processing in which a box filter is applied to the speckle reduced image to compute the average value of the pixels in a 3*3 window. This will lead to generate a multilooked image with smoothed and averaged version of the input. An additional median blur is applied to the multilooked image to help reduce the noise and enhance the image. As a final step normalisation is applied to the filtered image to ensure that the pre-processed image is within the standard scale suitable for visualisation and analysis. These preprocessing steps were defined as a function and used in ImageDataGenerator.

3.4.2 Data Augmentation

Using higher resolution in transfer learning with ResNet models tend to give better classification accuracy (Mahbod et. al, 2021). Instead of using the default image size of 224*224, the original image size of 400*400 is used as the input size in this research. After preprocessing the images data augmentation like horizontal flipping, shear range and zoom range was applied.

3.4.3 Train, validation, and test data

Five percent of the total images from each class were moved to a separate folder to test the model on unseen data. Remaining 95 percent of the images were used for training and validation phase. The validation split was defined as 0.1 or 10 percent of the previously divided training data. Separate data generators for training, validation and testing were used to get the data from respective directories.

3.5 Modelling

This section gives in depth detail about the steps involved in creating various deep learning models using Residual Network variants such as ResNet50, ResNet50V2 and ResNet101.

ResNet50

It is a variant of ResNet which consists of 50 layers excluding final fully connected layer. It consists of repeating blocks of layers with multiple convolution layers and identity shortcuts which enables the model to skip layers for learning residual functions. ResNet50 is widely used in image classification tasks. It is often used as a benchmark architecture.

ResNet50V2

This is an improved version of ResNet50 which includes minor changes in the architecture to incorporate training efficiency and generalization. By including pre activation structure it is helpful in dealing with gradient flow which in turn improves overall performance by being computationally efficient.

ResNet101

It's another variant of Residual network which is deeper than ResNet50 consisting of 101 layers without the last fully connected layer. These additional layers allow this model to capture complex hierarchical features from the image data, improving the ability of the model to capture more detailed understanding of the input. As in ResNet50, ResNet101 uses residual blocks with identity shortcuts.

3.5.1 Model building

Transfer learning is used because the pretrained models will be able to capture better features than the custom models. Models such as ResNet50, ResNet50V2, and ResNet101 utilises the weights which was trained on the larger ImageNet dataset. These pretrained models are used as the backbone and modified it for the specific requirement of binary classification. This modification includes addition of layers such as Flatten, Dense, Batch normalisation and sigmoid activation function in the final dense layer since it is a binary image classification problem. Since the ResNet model used as the backbone produces a multidimensional output containing spatial information is flattened using Flatten layer to convert the multidimensional tensor into one dimensional array. Which can then be fed to the fully connected layers. The dense layer with 512 neurons and ReLU activation function allows to learn complex patterns and representations from the data. The batch normalization layer after the dense layer is used since it can lead to faster convergence during training and contribute to better generalization. The final dense layer is used to meet the custom requirements of the classification. It creates a dense layer with neurons equal to the number of classes in the dataset. Since this research addresses a binary classification problem sigmoid activation function makes the values in range of [0,1] which in turn gives the probability for each class.

These modifications are kept same across all the models in the study to analyse the capability of each pre-trained models.

3.5.2 Model Training

The model is compiled with Adam optimizer with the default learning rate and sparse categorical cross entropy loss. Also, all models were trained for 30 epochs so that the performance of each model at the same epoch can be evaluated. Figure shows the training and validation loss of ResNet50, ResNet50V2, and ResNet101 up to 30 epochs.

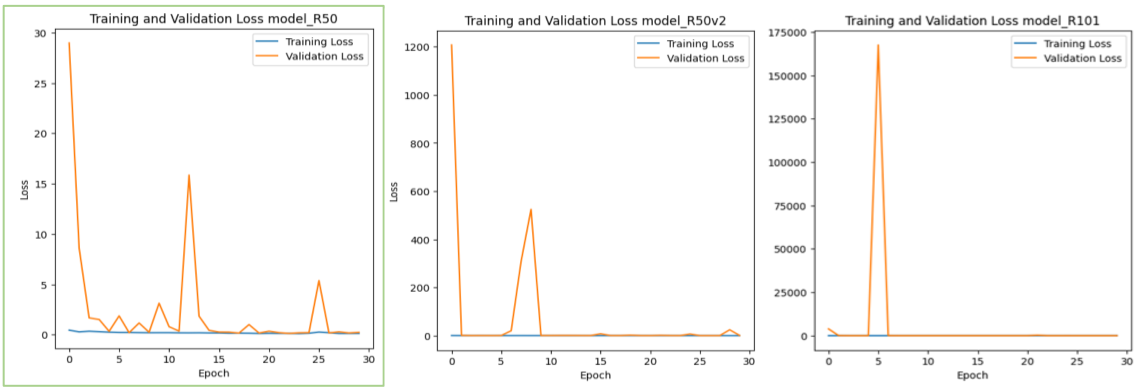

Figure 5. Training and Validation loss of ResNet50(left), ResNet50V2(centre), and ResNet101(right).

From Figure 5 we can see that the training loss for all model stays close to zero achieving a stable stage where weights and biases have the optimum values for predictions on training data especially for ResNet50 highlighted in green border. The highest training accuracy achieved is 96.13%. Whereas the validation loss tends to improve after each epoch with a highest value of 96.06%. The ModelCheckpoint callback is employed to save the best model based on validation accuracy during training.

3.6 Model Evaluation methods

Since the study addresses binary classification with class imbalance as identified from EDA, metrics like weighted Score, Confusion Matrix and ROC curve will be analysed.

4 Implementation

The implementation of the proposed approach is graphically depicted in Figure 6. The architectural diagram shows each step that was carried out to efficiently complete the proposed methodology. As depicted in Figure 6, The SAR image dataset used for this research was downloaded from the CSIRO data access portal. The data was used to perform EDA. Upon performing EDA, it was evident that the image dataset was imbalanced, and the pixel analysis showed that mean pixel value of the images containing oil like features is less than that of the images with non-oil like features. Since the class imbalance was identified during EDA helped to choose most suitable evaluation metrics like, F1-Score, confusion matrix and ROC curve. The data was pre-processed by applying calibration, speckle reduction, multilooking, filtering and normalization.

After preprocessing the entire image dataset, test data was moved to a new directory. The remaining dataset was used for training and validation. Training and validation dataset were augmented by applying horizontal flipping, shear range and zoom range. After data preparation, pretrained ResNet models such as ResNet50, ResNet50V2 and ResNet101 used for training and validation. With the help of pretrained model, time and resources required for building a model from scratch can be avoided. These pretrained models were trained on the larger imagenet dataset with weights set to the imagenet. Additional layers were added on top of these models to meet the binary classification requirement of this study.

ModelCheckPoint callback, loss, and optimizer were defined before compiling the model. After monitoring the validation accuracy, the best model will be saved to the specified location with the help of ModelCheckPoint. Sparse Categorical cross entropy was used since it is most suitable for the binary classification. Due to the adaptive learning rates and optimization efficiency Adam optimizer was used as the optimizer function. Then each compiled models were trained for 30 epochs.

After training, the model is evaluated on the test dataset. The classification report, confusion matrix, ROC curve was used to evaluate the performance since the dataset was biased. These evaluation metrics were analysed by visualising the results. These results where further used to identify feasibility of the proposed method. This process implementation architecture helped to effectively complete this study.

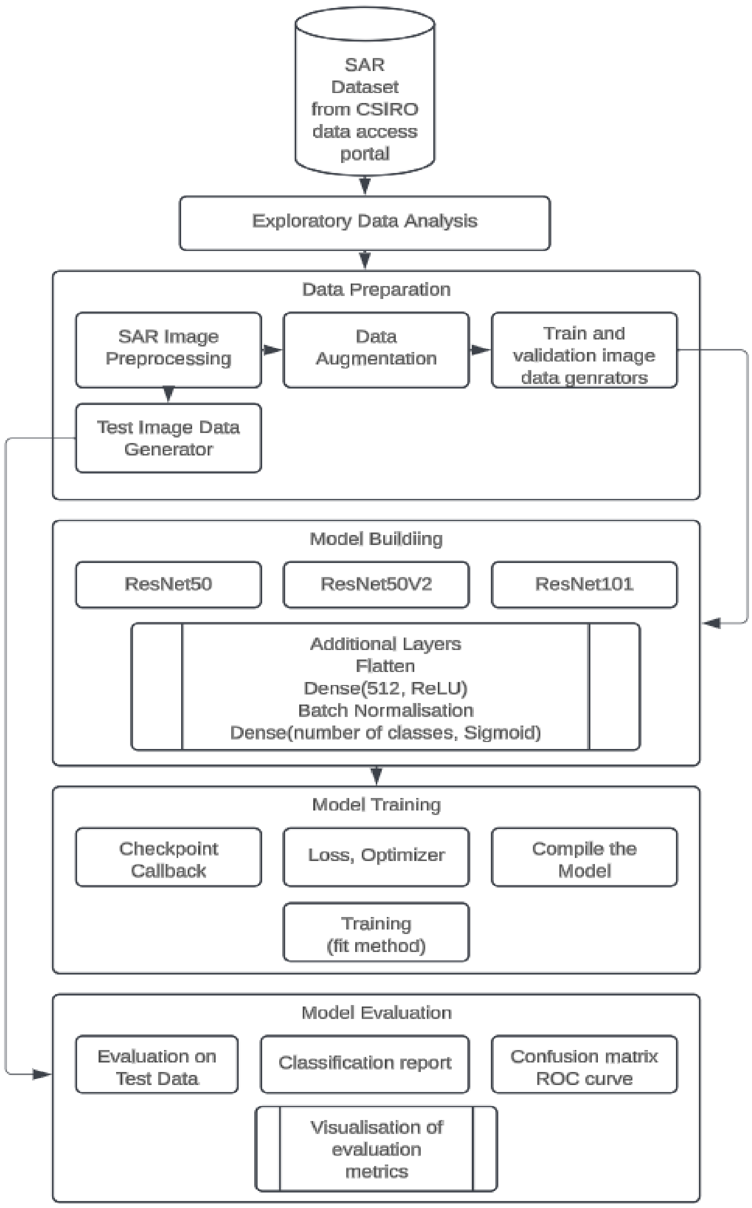

Figure 6. Architecture diagram of process implementation

5 Evaluation

Evaluating the trained models on the unseen test data can aid to understand the generalisability of these models. During training the models can show good performance but stumble upon unseen data. To check the generalisability of the model total of 281 images belonging to both classes were chosen randomly among which 186 images belongs to class 0 and 95 images belong to class 1. As mentioned in the research methodology the SAR image dataset was having class imbalance. To evaluate the performance of the models with class imbalance metrics such as weighted F-score, area under ROC curve and confusion matrix can be used. These methods can give better interpretation on the performance of the models.

5.1 Classification Report

Classification Report can be used to get the precision, recall, F1-score, and weighted average of these values. Along with the count of samples from each class which are subjected for testing. Table 2 gives the classification reports of each ResNet models used in this research for easy comparison. Number of samples from each class is same for all the models. Precision will allow us to understand the accuracy of positive predictions. That is how well the model was able to correctly classify class 0 and class 1. Recall helps to identify the model's ability to capture positive instances. Where as F1-score is the harmonic mean between precision and recall. The weighted average of precision, recall and F-score is an overall performance indicator while dealing with imbalanced datasets. |Support gives the number of class samples which is considered for the evaluation.

Model		Precision	Recall	F1-score	Support
ResNet50	**0**	0.97	0.95	0.96	186
	1	0.90	0.95	0.92	95
	Weighted Average	**0.95**	**0.95**	**0.95**	281
ResNet50V2	**0**	0.85	0.97	0.91	186
	1	0.91	0.67	0.78	95
	Weighted Average	0.87	0.87	0.86	281
ResNet101	**0**	0.89	0.96	0.92	186
	1	0.91	0.76	0.83	95
	Weighted Average	0.89	0.89	0.89	281

Table 2. Classification report of ResNet50, ResNet50V2 and ResNet101 on the test data

From Table 2 it is evident that the ResNet50 is giving better results, highlighted in bold. Compared to ResNet50V2 and ResNet101, ResNet50 was able to make better positive predictions and was able to capture all the positive instances.

Since the dataset was imbalanced the better way to understand the evaluation trade-off between the classes weighted average plays a crucial role.

In summary, with in a total test sample space of 281, in which 186 instances were belonging to class 0 and 95 instances belonging to class1, ResNet50 with a weighted Precision, recall and F1-score value of 0.95 outperforms all the other models, such as ResNet50V2 and ResNet101, used in this study.

5.2 Confusion Matrix

The figure 7 displays the confusion matrices used to assess the ability to accurately predict each class by various ResNet models. These confusion matrices can be compared with the confusion matrix got after using VGG19 in the base paper (Blondeau-Patissier, et al., 2023). These confusion matrixes show the trade-off between True Positives, True Negatives, False Positives and False Negatives.

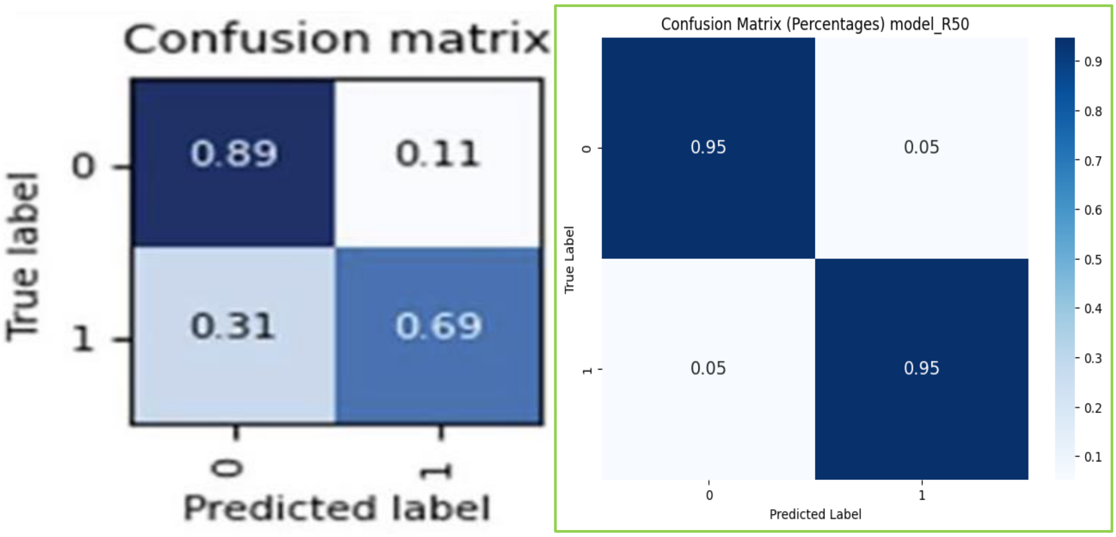

Figure 7. Confusion Matrix for VGG19(left) from (Blondeau-Patissier et al., 2023), ResNet50(right)

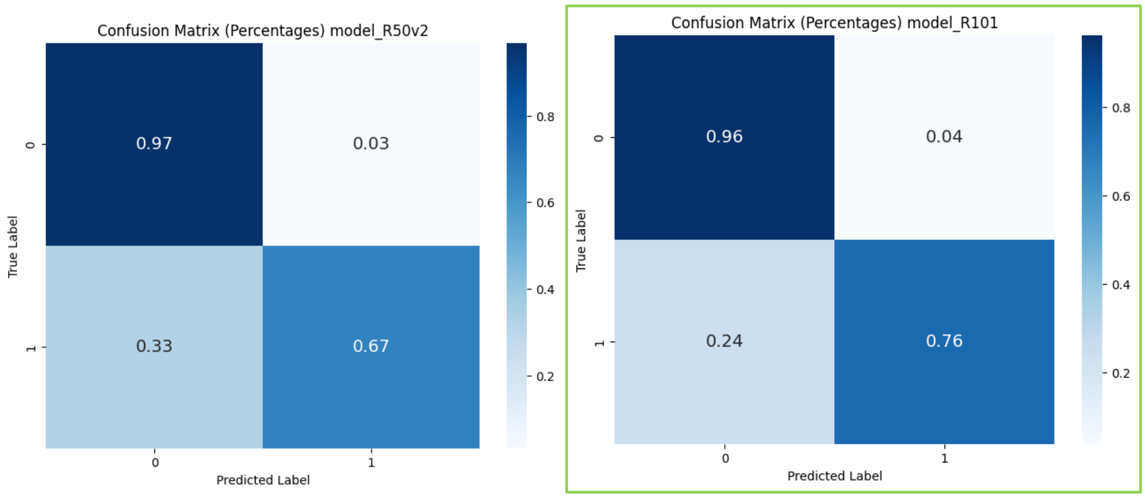

Figure 8. ResNet50V2 (left) and ResNet101(right)

As shown in Figure7 and Figure8, True positive and True negative classification percentage achieved by ResNet50 and ResNet101 models, highlighted in green border surpasses the classification percentages achieved by VGG19 model. ResNet50V2 model was able to surpass VGG19 in classifying the majority class but falls short in classifying the minority class.

5.3 ROC curve

According to the article discussed in Towards Data Science (Narkhede,2018), ROC curve is an important evaluation metrics to identify the performance of a classification model. It gives an idea about TPR (True Positive Rate) and FPR (False Positive Rate). If the majority area falls above the reference line near to 1 means that the model has good measure of separability. This can also help to understand the sensitivity and specificity of the classification model.

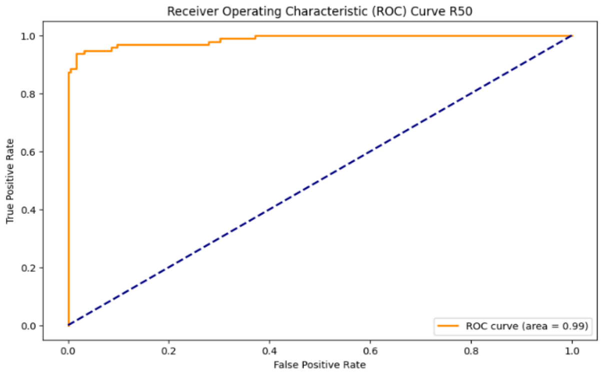

Figure 9. ROC curve of ResNet50

ROC curve area for ResNet50 is plotted in Figure 9 with an excellent value of 0.99 showing models outstanding capability to distinguish each class.

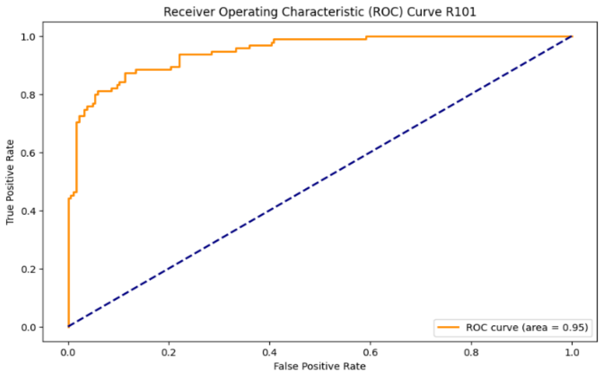

Figure 10. ROC curve of ResNet101

ROC curve area for ResNet101 is plotted in Figure 10 with a promising value of 0.95 showing models capability to distinguish each class.

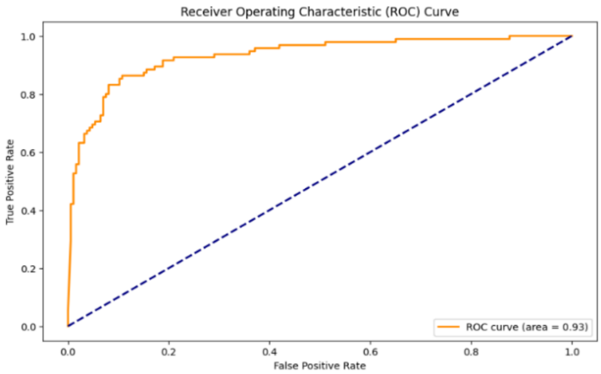

Figure 11. ROC curve of ResNet50V2

ROC curve area for ResNet50V2 is plotted in Figure 11 with a comparatively lesser value than other ResNet models with area of 0.93 showing models capability to differentiate each class.

The Figure 9, 10 and 11 shows that all the models give ROC curve closer to 1 which means the models have excellent discriminative power. The ROC curve area of ResNet50 is 0.99, Reset101 is 0.95 and ResNet50V2 is 0.93. All the models used in this study performed well on classifying binary classes. ResNet50 surpasses other models with higher TPR value closer to 1 (0.99). A value of 0.99 suggests that the model is making effective predictions across a wide range of threshold settings. It means that there is 99% chance where model will be able to distinguish between the classes. Also, it indicates that the model is achieving great balance between sensitivity and specificity.

6 Discussion

The ResNet models used for this research gave promising results with lesser misclassification than the VGG19(Blondeau-Patissier et al., 2023). The detailed deduction of results is explained in this section.

6.1 Classification Performance

The classification report in Table 2 provides a detailed breakdown of precision, recall, and F1-score for each ResNet model. ResNet50 outperforms ResNet50V2 and ResNet101 with a weighted F-score, precision, and recall of 0.95. This superior performance is particularly crucial in the context of the imbalanced dataset, where ResNet50 demonstrates its ability to handle both classes effectively. Also, the weighted Fscore of the ResNet50 surpassed VGG19 by 0.05 resulting in 0.95 compared to 0.90.

6.2 Confusion Matrix

As shown by the confusion matrix in Figure 7 and Figure 8, Even though ResNet50 and ResNet101 gave better results than VGG19, the ResNet50 model was able to classify the oil and non-oil like images well with only 5% percent misclassification in both classes compared to 11% False positives and 31% false negatives while using VGG19.

6.3 ROC curve analysis

The ROC curve analysis (Figure 9, Figure 10, and Figure 11) further confirms the excellent discriminative power of the ResNet models. ResNet50 stands out with an ROC curve area of 0.99, indicating its ability to distinguish between classes effectively. This high value suggests a great balance between sensitivity and specificity, suggesting the robustness of ResNet50 in making accurate predictions.

6.4 Model Comparison

Table 3 summarizes the comparison between ResNet50 and VGG19, showcasing superior performance in terms of training accuracy, validation accuracy, test accuracy, and F1-score. The ResNet50 model demonstrates consistent improvements across all metrics, emphasizing its effectiveness in addressing the challenges posed by SAR image classification for oil spill detection.

Metrics	ResNet50	VGG19
Training Accuracy	**0.9613**	0.95
Validation accuracy	**0.9606**	0.90
Test accuracy	**0.9537**	0.90
Fscore	**0.95**	0.90

Table 3. Comparison between ResNet50 VS VGG19

As shown in Table 3, the proposed approach in this study using ResNet50 surpasses the VGG19 model (Blondeau-Patissier et al., 2023). The usage of the original image size, optimal epochs, data augmentation and SAR image preprocessing played a vital role in making better classification.

6.5 Limitations of the work

The proposed approach shows improvements whereas the imbalanced dataset poses a challenge, with weighted metrics giving out a balanced assessment. Handling imbalanced dataset using new and innovative methods should have been considered. In this study ResNet models were used more models need to be tried and performance should have been analysed. Also, the study focuses on GBR marine park so it may lack generalisation while handling datasets from different regions.

7 Conclusion and Future Work

Marine contamination due to oil spills becomes an important concern on marine habitat. Utilising suitable Deep Learning methods are inevitable since the vast area of ocean cannot be monitored manually. Deep Learning methods can give generalisability over any other approaches. This research compared various ResNet models to find out the effectiveness in classifying SAR images containing oil like and non-oil like features with the VGG19 model proposed. All the ResNet model performed well on unseen dataset, but the ResNet50 surpassed ResNet50V2, ResNet101 and VGG19 with highest weighted Fscore of 0.95 and ROC curve area of 0.99. The usage of the original image size, optimal epochs, data augmentation and SAR image preprocessing would have helped to generate better classification results.

This research was able to effectively perform binary classification on oil like and non-oil like features. The imbalanced nature of the dataset calls for considering new and innovative methods for over sampling or under sampling (Tyagi and Mittal, 2020) to handle the issue. The non-oil like class used in this research contains clear sea, biogenic slicks, and other look-alikes. Suitable methods like weakly supervised segmentation (Luo et al., 2021) to perform semantic segmentation on this non-annotated dataset should be explored which can help to classify the features more accurately since manual annotation methods can be expensive and time consuming.

References

Al-Ruzouq, R., Gibril, M.B.A., Shanableh, A., Kais, A., Hamed, O., Al-Mansoori, S. and Khalil, M.A., 2020. Sensors, features, and machine learning for oil spill detection and monitoring: A review. Remote Sensing, 12(20), p.3338.

Blondeau-Patissier, David; Schroeder, Thomas; Diakogiannis, Foivos; Li, Zhibin (2022): CSIRO Sentinel-1 SAR image dataset of oil- and non-oil features for machine learning (Deep Learning). v1. CSIRO. Data Collection. https://doi.org/10.25919/4v55-dn16

Blondeau-Patissier, D., Schroeder, T., Suresh, G., Li, Z., Diakogiannis, F.I., Irving, P., Witte, C. and Steven, A.D., 2023. Detection of marine oil-like features in Sentinel-1 SAR images by supplementary use of deep learning and empirical methods: Performance assessment for the Great Barrier Reef marine park. Marine Pollution Bulletin, 188, p.114598.

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

Huang, X., Zhang, B., Perrie, W., Lu, Y. and Wang, C., 2022. A novel deep learning method for marine oil spill detection from satellite synthetic aperture radar imagery. Marine Pollution Bulletin, 179, p.113666.

Jatiault, R., Dhont, D., Loncke, L. and Dubucq, D., 2017. Monitoring of natural oil seepage in the Lower Congo Basin using SAR observations. Remote Sensing of Environment, 191, pp.258-272.

Krestenitis, M., Orfanidis, G., Ioannidis, K., Avgerinakis, K., Vrochidis, S. and Kompatsiaris, I., 2019. Oil spill identification from satellite images using deep neural networks. Remote Sensing, 11(15), p.1762.

Luo, W., Yang, M. and Zheng, W., 2021. Weakly-supervised semantic segmentation with saliency and incremental supervision updating. Pattern Recognition, 115, p.107858.

Mahbod, A., Schaefer, G., Wang, C., Ecker, R., Dorffner, G. and Ellinger, I., 2021, January. Investigating and exploiting image resolution for transfer learning-based skin lesion classification. In 2020 25th international conference on pattern recognition (ICPR) (pp. 4047-4053). IEEE.

Mascarenhas, S. and Agarwal, M., 2021, November. A comparison between VGG16, VGG19 and ResNet50 architecture frameworks for Image Classification. In 2021 International conference on disruptive technologies for multi-disciplinary research and applications (CENTCON) (Vol. 1, pp. 96-99). IEEE.

Mdakane, L.W. and Kleynhans, W., 2020. Feature selection and classification of oil spill from vessels using Sentinel-1 wide–swath synthetic aperture radar data. IEEE Geoscience and Remote Sensing Letters, 19, pp.1-5.

Narkhede, S. (2018) Understanding AUC - roc curve, Towards Data Science. Available at: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5 (Accessed: 30 November 2023).

Shu, Y., Li, J., Yousif, H. and Gomes, G., 2010. Dark-spot detection from SAR intensity imagery with spatial density thresholding for oil-spill monitoring. Remote Sensing of Environment, 114(9), pp.2026-2035.

Suresh, G., Melsheimer, C., MacDonald, I. R., Notholt, J., & Bohrmann, G. (2017). Application of the automatic seep location estimator (ASLE) with the use of contextual information for estimating offshore oil seeps. Remote Sensing Applications: Society and Environment, 5, 16-26. https://doi.org/10.1016/j.rsase.2016.11.005

Tyagi, S. and Mittal, S., 2020. Sampling approaches for imbalanced data classification problem in machine learning. In Proceedings of ICRIC 2019: Recent Innovations in Computing (pp. 209-221). Springer International Publishing.

Yuan, Q., Shen, H., Li, T., Li, Z., Li, S., Jiang, Y., Xu, H., Tan, W., Yang, Q., Wang, J. and Gao, J., 2020. Deep learning in environmental remote sensing: Achievements and challenges. Remote Sensing of Environment, 241, p.111716.

Zhang, B., Matchinski, E.J., Chen, B., Ye, X., Jing, L. and Lee, K., 2019. Marine oil spills—Oil pollution, sources and effects. In World seas: an environmental evaluation (pp. 391-406). Academic Press.

Zheng, X. and Chen, T., 2021. High spatial resolution remote sensing image segmentation based on the multiclassification model and the binary classification model. Neural Computing and Applications, pp.1-8.

Zhou, H. and Peng, C., 2018, July. Oil spills identification in SAR image using mRMR and SVM model. In 2018 5th International Conference on Information Science and Control Engineering (ICISCE) (pp. 355-359). IEEE.