Water Pollution Detection Through Hyperspectral Images

Project Team

Shouzab Khan	p17-6101
Zunera Bukhari	p17-6052
Shahab Aslam Paracha	p17-6132

Session 2017-2021

Supervised by

Muhammad Amin

Department of Computer Science

National University of Computer and Emerging Sciences
Peshawar, Pakistan

January, 2022

Student's Declaration

We declare that this project titled "*Water Pollution Detection Through Hyperspectral Images*", submitted as requirement for the award of degree of Bachelors in Computer Science, does not contain any material previously submitted for a degree in any university; and that to the best of our knowledge, it does not contain any materials previously published or written by another person except where due reference is made in the text.

We understand that the management of Department of Computer Science, National University of Computer and Emerging Sciences, has a zero tolerance policy towards plagiarism. Therefore, We, as authors of the above-mentioned thesis, solemnly declare that no portion of our thesis has been plagiarized and any material used in the thesis from other sources is properly referenced.

We further understand that if we are found guilty of any form of plagiarism in the thesis work even after graduation, the University reserves the right to revoke our BS degree.

Shouzab Khan Signature: ________________

Zunera Bukhari Signature: ________________

Shahab Aslam Paracha Signature: ________________

____

Verified by Plagiarism Cell Officer
Dated:

Certificate of Approval

The Department of Computer Science, National University of Computer and Emerging Sciences, accepts this thesis titled *Water Pollution Detection Through Hyperspectral Images*, submitted by Shouzab Khan (p17-6101), Zunera Bukhari (p17-6052), and Shahab Aslam Paracha (p17-6132), in its current form, and it is satisfying the dissertation requirements for the award of Bachelors Degree in Computer Science.

Supervisor

Muhammad Amin Signature: ____________________

Mashal Khan

FYP Coordinator
National University of Computer and Emerging Sciences, Peshawar

Dr. Muhammad Hafeez

HoD of Department of Computer Science
National University of Computer and Emerging Sciences

Acknowledgements

We are very thankful to almighty Allah for giving us a lot of courage to choose such an immense project and for progressing in it with great dignity and grace. All the group members played significant role in the progress of this project. We would like to grab the opportunity to thank Muhammad Amin who motivated us for selection of this project. They encouraged us and showed us the way the project can be implemented. They have been really kind to us since the beginning of our project. Despite of their very busy schedule as lecturer they were always available to us and gave us time whenever we wanted to meet him. Their guidance and experience proved very helpful to us in order to make progress in our project. We are highly thankful to all of our teachers who had been guiding us throughout our project work. Their knowledge, guidance and training enabled us to carry out this development work more efficiently. We would like to thank all our friends and colleagues for giving us useful suggestions about increasing the quality of our project. Their suggestions and motivation helped us a great deal.

Shouzab Khan
Zunera Bukhari
Shahab Aslam Paracha

Abstract

Water is life. Water, of being great importance, espacially inland waters, requires devotion and efforts for it to be monitered well and for a better quality of it to be provided. We are providing means of doing so.

We analyze the relationship of water using "Hyperspectral Remote Sensing". We can also use hyperspectral imagery to evaluate overall quality of water.. Hyperspectral Remote Sensing has possible means of presenting you details of the contamination rapidly and inexpensively.

A self-adapting selection method of multiple artificial neural networks (ANN) is proposed to experimentally predict water quality parameters such like "phosphorus", "nitrogen", "biochemical oxygen demand (BOD)", "chemical oxygen demand (COD)", and "chlorophyll-a" using Hyperspectral Remote Sensing and earth measured water qualitative data.

An ongoing scientific issue is evaluating the absorption spectra of the water column using hyperspectral images. Convolutional neural networks (CNN) have now become a popular strategy for classification and regression on huge datasets due to recent breakthroughs in deep learning. To estimate and therefore monitor several indices for water quality, such as "chlorophyll-a" and "nitrogen," we propose a combination of hyperspectral data and machine learning techniques. Hyperspectral Images have also been shown to improve in the classification of "chlorophyll-a." **Keywords:** Water Quality, Chlorophyll-a, nitrogen, hyper-spectral images, machine learning, regression.

Contents

1 Preliminaries and Introduction **1**

1.1 INTRODUCTION 1

1.2 Background 3

1.3 Why Hyperspectral Images? 3

1.4 Motivation 4

1.5 Problem Statement 5

1.6 Objectives 5

2 Review of Literature **7**

2.1 PAPERS 7

2.2 Some Machine Learning Methods 7

2.2.1 Dimensionality Reduction 8

2.2.2 Principal Component Analysis(PCA) 9

2.3 Classification Algorithms 11

2.3.1 Support Vector Machine 11

3 System Analysis Design **13**

3.1 WORK FLOW 13

3.2 Use Case 13

3.3 Use Case Diagram 14

3.4 Activity Diagram 14

4 Methodology **15**

4.1 Data Collection 15

4.2 Data Preprocessing 15

4.2.1 Reading Dataset 16

4.2.2 Extracting Pixels 16

4.3 Exploratory Data Analysis 17

4.3.1 Interactive Visualizations 17

4.3.1.1 Bar Plot 17

4.3.1.2 Pair plot 18

4.3.1.3 3D Scatter Plot 18

4.3.1.4 Area Plot 19

4.3.2 Box Plot 19

4.4 Dimensionality Reduction(DR) 20

4.4.1 Feature Extraction 20

4.4.2 Feature Selection 20

4.5 Principal Component Analysis(PCA) 20

4.5.1 Mathematics Behind PCA 21

4.5.2 How many dimensions should we reduce? 21

4.6 Classification Algorithms 22

4.6.1 Support Vector Machine(SVM) 23

5 Results **25**

5.1 VISUALIZING AND FEATURE EXTRACTION 25

5.1.1 BANDS 25

5.1.2 GROUND TRUTH 25

5.1.3 SPECTRAL SIGNATURE 26

5.1.4 CUMULATIVE VARIANCE 26

5.1.5 BANDS AFTER PCA 26

5.1.6 CONFUSION MATRIX 27

5.1.7 CLASSIFICATION REPORT 27

5.1.8 MODEL ACCURACY 27

5.1.9 FINAL BAND 29

5.2 EXPERIMENTAL RESULTS 29

6 Discussion **31**

7 Conclusion and Future Work 33

References 38

List of Figures

1.1 Hyperspectral Image Spectrum 4

1.2 Hyperspectral Image Graph 4

2.1 Summary for all literature review 8

2.2 Standardization 9

2.3 Covariance matrix computation 10

2.4 Compute the eigenvectors and eigenvalues 10

2.5 Recast the data along the principal components axes 10

2.6 Different Hyperplane 12

2.7 Different Suppot Vectors 12

3.1 Work Flow 14

3.2 Use Case 14

4.1 Date in dataset 16

4.2 Bar Plot 17

4.3 Pair Plot 18

4.4 Scattered Plot 18

4.5 Area Plot 19

4.6 Box PLot 19

4.7 PCA 21

4.8 Extracted 4 band after DR 22

5.1 Band Extracted From Our Dataset 26

5.2 Ground Truth Extracted From Our Dataset 26

5.3 Spectral Signature of Band 27

5.4 Cumulative Variance Between Band 27

5.5 Extracted Band After PCA 27

5.6 Performance Of Model 28

5.7 Quality Of Prediction 29

5.8 Final Band Extracted After PCA 30

List of Tables

Chapter 1

Preliminaries and Introduction

The chapter begins with an introduction to the fundamental ideas of water contamination before moving on to a discussion of the data that will be used for this project. It also goes over the Hyperspectral Images that we utilised.

1.1 INTRODUCTION

The particles in the water are dangerous to the environment and many other factors. Impurities in the atmosphere, such as "metals," "algae," "chlorophyll-a," and other dust particles, pollute the environment and endanger aquatic creatures, plants, and human health. "Water resources are used to generate economic growth in several countries. Water provides food, money, and a way of life for many creatures. The essential water quality metrics are "phosphorus", "nitrogen", "biochemical oxygen demand (BOD)", "chemical oxygen demand (COD)", and "chlorophyll-a" (Chl-a). The measure of "chlorophyll-a" is an important indicator of water quality, nutritional state, and organic contamination. According to Cheng et al. (1) , such data is useful for regulating water quality and monitoring water contamination (2013). Plants, such as "phytoplanktonic algae", have "chlorophyll-a," the primary green pigment. Photosynthetic plankton biomass is estimated using "chlorophyll-a" concentrations in coastal or marine waters ". Excess nitrogen and phosphorus in water are responsible for high levels of nutrients in the water, which can come from various sources such as crop fertilizers, animal dung, and industrial waste. "Furthermore, dense algae development prevents other aquatic species from receiving the light they require, low oxygen kills seagrass, fish, crabs, oysters, and other sea creatures, according to study. Low levels of "chlorophyll-a" indicate "chlorophyll-a" levels proxy for nutrient availability), while high levels indicate poor water quality. Traditional monitoring methods involve taking point-based measurements of water quality parameters in the lab. These calculations are exact and provide a comprehensive depth profile at a specific location. They are covering ample water are covering spatially confined, costly, and time-consuming. Advanced advancements in hyperspectral have opened up new options for unique data collection throughout the last decade. Hyperspectral cameras have a high spectral resolution and analyze water qualms throughout a wavelength range of "450 to 950 nano meters". In general, Hyperspectral sensors detects the reflectance of water components. The reflectance of the quality metrics are referred to as their spectral signatures in the following. The gases emissions from in following factories and businesses significantly impact the environment. Even though government regulatory organisations in some countries organizations precise emission thresholds for particles, they cannot prevent the harmful effects of these particles in water". Using hyperspectral images to extract seawater properties has been the target of several research, including one to Using hyperspectral imagery to improve the process of monitoring seawater parameters(2). "The use of the Linear Matrix Inversion (MMI) approach to estimate seawater parameters has substantial limits, notably in coastal locations where water quality is nonlinear, as compared to the open sea or ocean, where the authors obtained an improvement over multispectral images.".

1.2 Background

“In recent years, random, artificial sampling has been used to detect water pollution, in which laboratory staff gather water samples at random for different testing in order to tally up the overall quantity of contamination. To significantly forecast water indices features, empirical approaches such as "Multispectral Index Analysis", Semi-analytic Methods such as "Hyperspectral Index analysis" and Deep Learning Methods such as "Artificial Neural Network (ANN)" analysis are frequently utilized(3). Hyperspectral remote sensing has become widely used in forecasting the atmosphere, soil, and water, thanks to the rapid advancement of computer science and remote sensing. "Chl-a", "total suspended solids" and "turbidity levels" are determined using spectral indices generated from hyperspectral or multispectral data. Because they have low computational costs, take little time and water quality parameters such as "nitrogen," "phosphorus," "BOD," "COD," and "Chl-a" are all predicted using spectral indices based on spectral images."

1.3 Why Hyperspectral Images?

"In remote sensing, hyperspectral remote sensors are often used to monitor the earth’s surface with high spectral resolution. The HSI typically has hundreds of bands compared to standard RGB photos. Hyperspectral images are used in crop analysis, geological mapping, mineral exploration, defense research, urban investigation, military surveillance, and other applications (HSI)".According to syam at el. (4) bare photos do not convey much information about the pixels and bands. As a result, hyperspectral photographs are employed, which provide more information per pixel. The spectral bands of Hyperspectral photographs include helpful information for identifying various materials. Hyperspectral remote sensors are commonly employed in remote sensing for monitoring the earth’s surface. Bands are continuous and not restricted to the visible spectrum—the RGB camera’s shortcoming compared to hyperspectral cameras. The "Specim FX10" hyperspectral camera captures the full spectral signature, allowing it to correctly quantify the differences between almonds and their shells independent of their physical qualities.

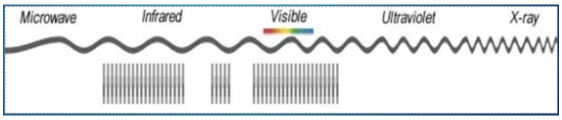


Figure 1.1: Hyperspectral Image Spectrum

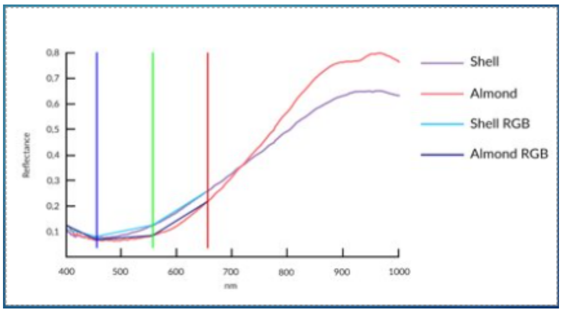


Figure 1.2: Hyperspectral Image Graph

Example of Hyperspectral images are given in figures 1.1, 1.2

1.4 Motivation

Water contamination has risen throughout time, posing a risk to human health. Water pollution detection in big bodies of water is a simple task. The ineffective detection of water contamination is posing a problem. Water contamination causes a variety of health issues. Among all kinds of water contamination, "nitrogen" and "chlorophyll-a" are the most overlooked and dangerous. "Remote sensing methods provide spatial and temporal perspectives on surface water quality parameters that are not readily available from in-situ measurements, allowing for effective and efficient landscape monitoring, as well as the identification and quantification of water quality parameters and concerns." The concept that inspired us to create this project is for users to be warned about the degree of water pollution before using specific water for personal use and get necessary health precautions based on the level of water pollution using appropriate pictures taken by hyperspectral satellites. It will also be the cost-efficient approach to analyze water in real-time. It will also be a cost-effective method of analyzing water in real-time. This initiative will be beneficial not just to humans and aquatic life but also to agriculture. "Nitrogen" and "chlorophyll-a" can also be detected, tracked, and forecasted.

1.5 Problem Statement

Because one never knows when water may become polluted, collecting many water samples for testing can be time-consuming. The dilemma is what to do if water becomes contaminated since Ammonia, lead, PH, or bacteria is hugely dangerous and cannot be seen with the human eye. In this case, we will identify these particles in the water and assess the water quality index in a given region.

1.6 Objectives

1. First of all our main objective was to find a suitable Dataset for our project on which we can relay.
2. After acquiring dataset then we had to visualize "Hyperspectral Images".
3. Preprocessing is another main step towards our project from before starting the main work.
4. Model Training is the step when we train the model on the given data.
5. Detection of "Nitrogen" and "Chlorophyll a" is the final step of our project when impurities will be detected from the dataset.

Chapter 2

Review of Literature

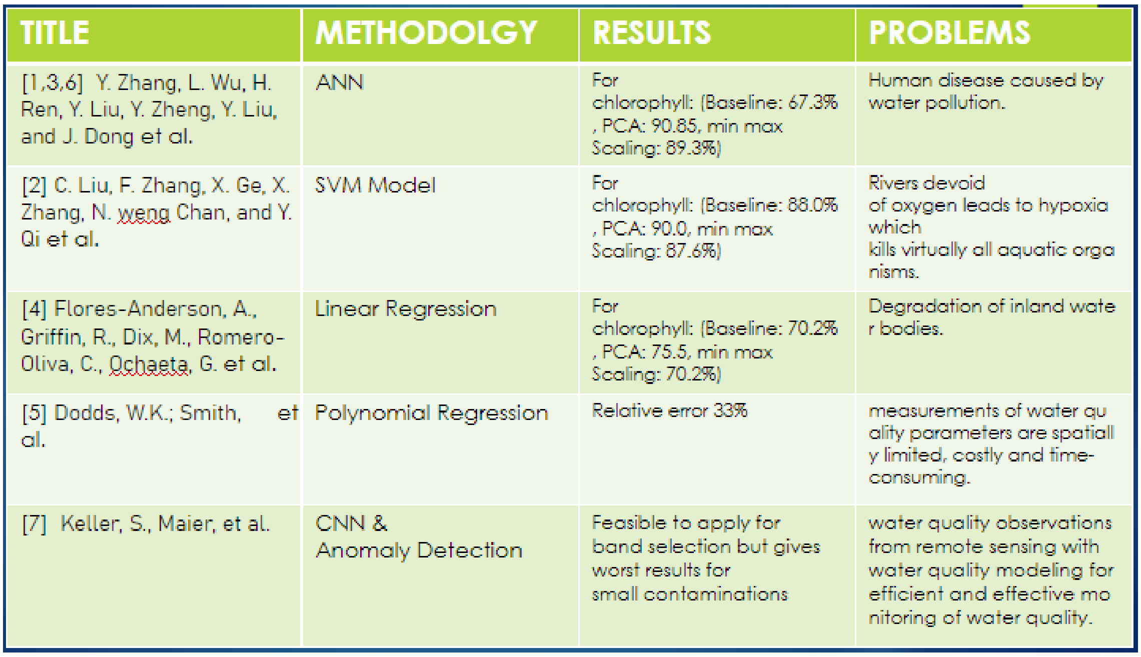
We have studied several publications and journals on various technologies and strategies for identifying water contamination using hyperspectral pictures. We have summarised the entire papers we had read for the understanding of our project.

2.1 PAPERS

We had read up to 20-25 papers for each and every concept related to our project. For each related topic we had done approx 4-5 papers for clearing our concepts and making the project work.
Following are the related papers for our project from which we had gone through:
(5) (6) (7) (8) (9)
(10) (11) (12) (13) (14)
(15) (16) (17) (18) (19) (20) (21) (22) (23) (24)
(25) (4) (3) (26) (2) (1)

2.2 Some Machine Learning Methods

Several machine learning models were utilized to identify contaminants in water using hyperspectral pictures. Many studies employ various regression and deep learning meth-



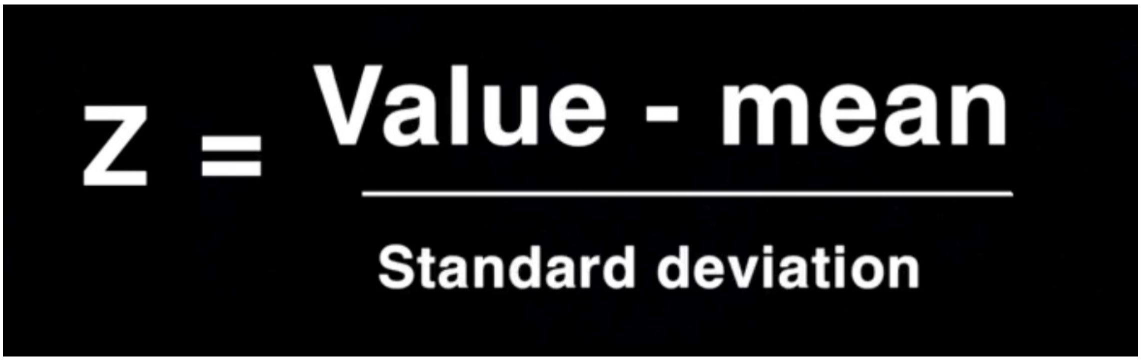
TITLE	METHODOLGY	RESULTS	PROBLEMS
[1,3,6] Y. Zhang, L. Wu, H. Ren, Y. Liu, Y. Zheng, Y. Liu, and J. Dong et al.	ANN	For chlorophyll: (Baseline: 67.3% , PCA: 90.85, min max Scaling: 89.3%)	Human disease caused by water pollution.
[2] C. Liu, F. Zhang, X. Ge, X. Zhang, N. weng Chan, and Y. Qi et al.	SVM Model	For chlorophyll: (Baseline: 88.0% , PCA: 90.0, min max Scaling: 87.6%)	Rivers devoid of oxygen leads to hypoxia which kills virtually all aquatic organisms.
[4] Flores-Anderson, A., Griffin, R., Dix, M., Romero-Oliva, C., Ochaeta, G. et al.	Linear Regression	For chlorophyll: (Baseline: 70.2% , PCA: 75.5, min max Scaling: 70.2%)	Degradation of inland water bodies.
[5] Dodds, W.K.; Smith, et al.	Polynomial Regression	Relative error 33%	measurements of water quality parameters are spatially limited, costly and time-consuming.
[7] Keller, S., Maier, et al.	CNN & Anomaly Detection	Feasible to apply for band selection but gives worst results for small contaminations	water quality observations from remote sensing with water quality modeling for efficient and effective monitoring of water quality.

Figure 2.1: Summary for all literature review

ods. The majority of the work on the following methodologies is summarised in figure 2.1

2.2.1 Dimensionality Reduction

"Dimensionality reduction has become an essential feature of machine learning due to the existence of a significant number of bands in the data. Dimensionality reduction has grown in popularity as a means of improving the accuracy of pixel categorization in hyperspectral images"(4). PCA, ICA, and LDA are some common dimensionality reduction approaches.



$$Z = \frac{\text{Value - mean}}{\text{Standard deviation}}$$

Figure 2.2: Standardization

“Dimensionality Reduction can be done in two types”:

1: Feature Selection

2: Feature Extraction

“Feature Selection”: “is the process of selecting dataset properties that contribute to machine learning tasks like classification and clustering. Several methods, such as correlation analysis, univariate analysis, and so on, can be used to do this”.

“Feature Extraction”: “Feature Extraction is the process of identifying new features by selecting and/or combining existing characteristics to reduce the feature space while still properly and comprehensively characterising the data set”.

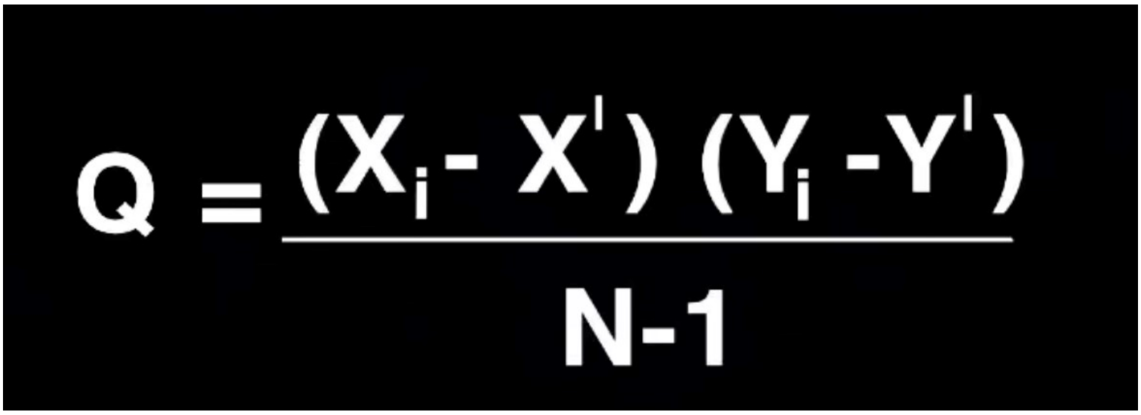
2.2.2 Principal Component Analysis(PCA)

PCA is a one of popular technique for Dimensionality Reduction. "Providing uncorrelated features with the greatest variance increases interpretability by minimizing information loss. It is challenging to analyze the "Pavia University dataset" in many dimensions. As a consequence, using Principal Component Analysis (PCA), the data dimensions are reduced to 3D.".

PCA has the following steps :

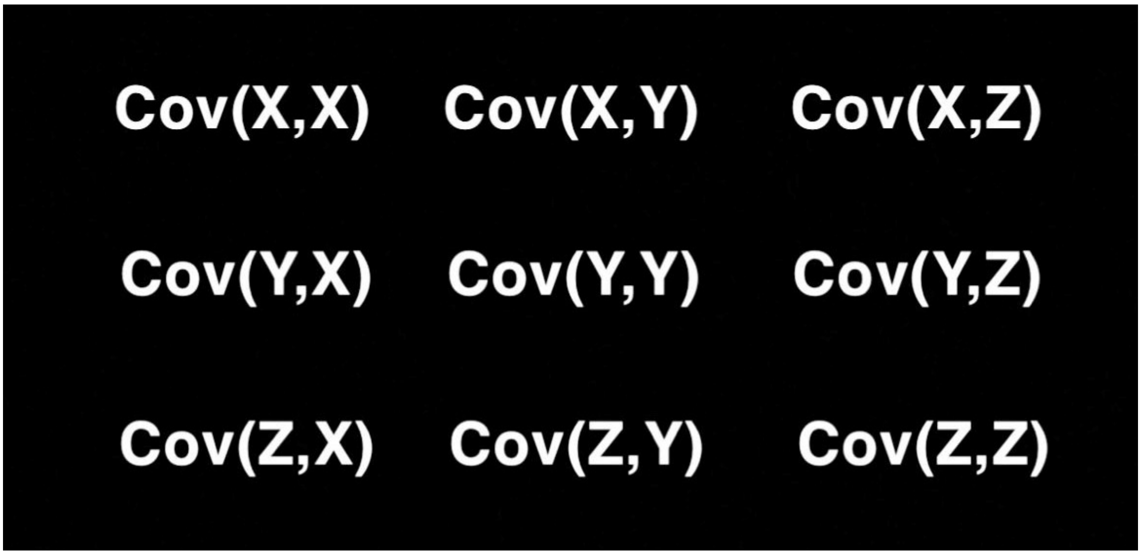
1. “Standardization”. 2.2

2. “Co-variance matrix computation”. 2.3



$$Q = \frac{(X_i - X^{'})\,(Y_i - Y^{'})}{N-1}$$

Figure 2.3: Covariance matrix computation



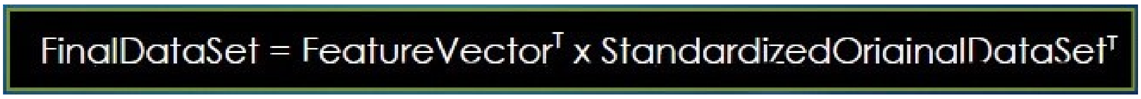
Cov(X,X)	Cov(X,Y)	Cov(X,Z)
Cov(Y,X)	Cov(Y,Y)	Cov(Y,Z)
Cov(Z,X)	Cov(Z,Y)	Cov(Z,Z)

Figure 2.4: Compute the eigenvectors and eigenvalues

3. “Compute the eigenvectors and eigenvalues of the co-variance matrix to identify the principal components”. 2.4

4. “Create a feature vector”.

5. “Recast the data along the principal components axes”. 2.5



$$\text{FinalDataSet} = \text{FeatureVector}^{T} \times \text{StandardizedOriginalDataSet}^{T}$$

Figure 2.5: Recast the data along the principal components axes

2.3 Classification Algorithms

“Classification refers to a predictive modeling problem where a class label is predicted for the given input data”. The classification can be divided as :

1. "Predictive Modeling Classification"
2. "Binary Classification"
3. "Multi-Class Classification"
4. "Multi-Label Classification"
5. "Imbalanced Classification"

Multi-Class Classification are the big issues which are been faced now adays. For the classification of hyperspectral images, many classification methods are utilised, including as:

1. "K-Nearest Neighbor"
2. "Support Vector Machine"
3. "Convolutional Neural Networks"

In this research paper, The "Support Vector Machine(SVM)” will be used to classify the "Hyperspectral Image".

2.3.1 Support Vector Machine

“The Support Vector Machine is a supervised classification method that optimizes the data-hyperplane margin. Several kernel functions are used to protect the data into higher dimensions”(4).

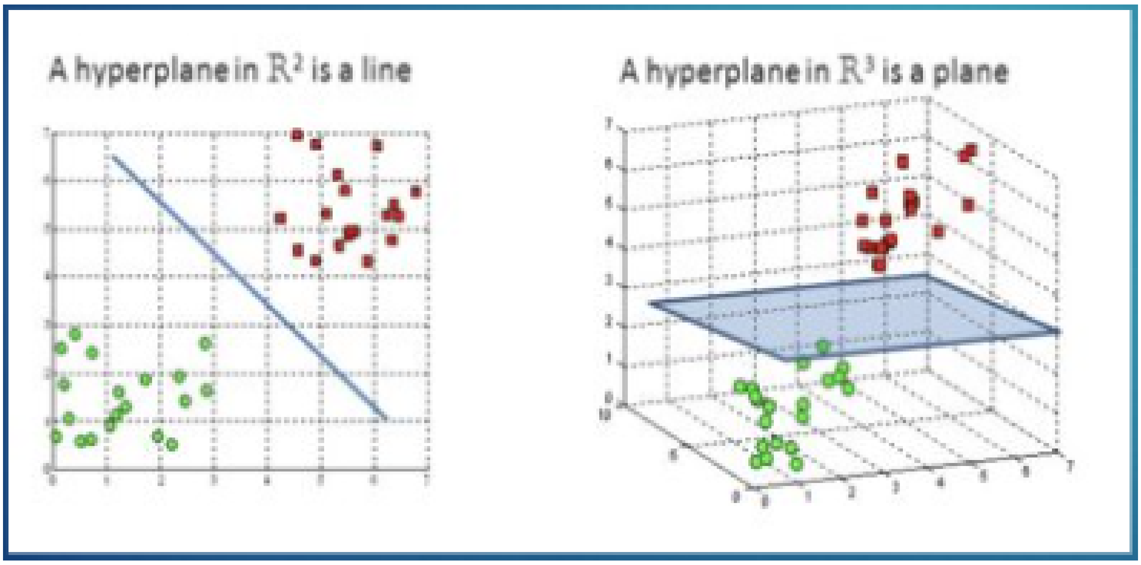


Figure 2.6: Different Hyperplane
(4)

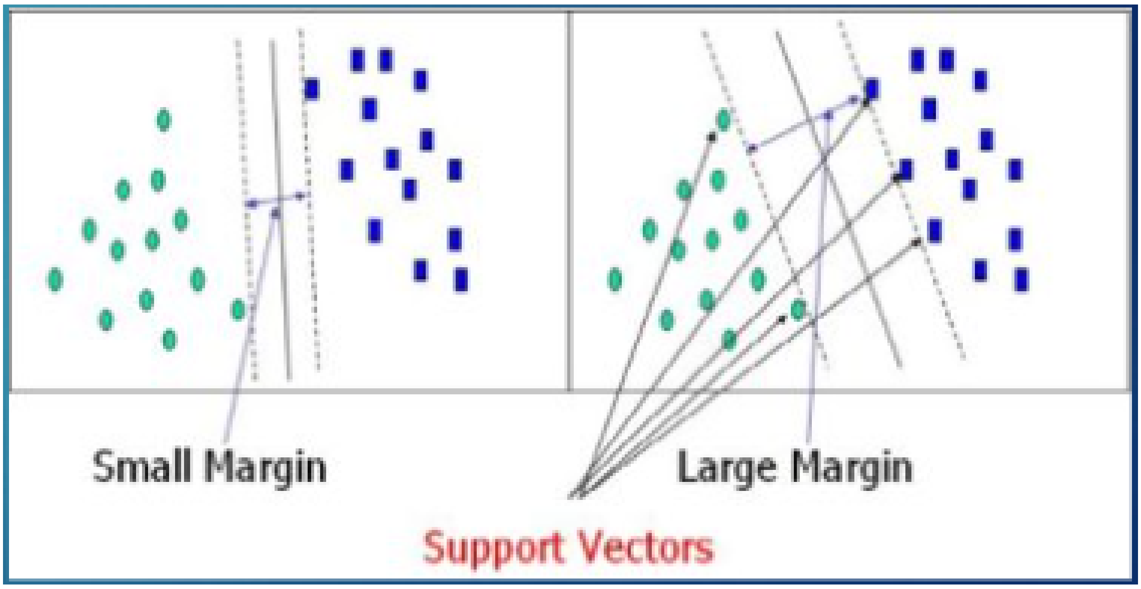


Figure 2.7: Different Suppot Vectors
(4)

Chapter 3

System Analysis Design

This chapter will analyze the system by creating several diagrams to evaluate how the system would behave when tested in the actual world. We create Unified Modeling Language (UML) diagrams to graphically depict the system and its essential properties to comprehend the system better.

3.1 WORK FLOW

Hyper-spectral satellite images were used to capture remote sensing spectra, which were then compared to measurements of "Chlorophyll-a" concentration. We'll go through how the measurements were taken, how satellite data was processed, and how the algorithm was designed and tested in the sections below. See figure 3.1

3.2 Use Case

One use for the Water pollution detection system is to deploy it in an environment to see if specific contamination may be detected or not. This use case may be applied to the environment, Defense sector, agriculture, Marine life and many more.

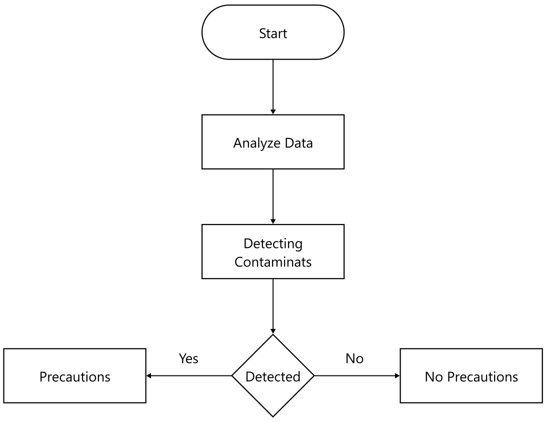


Figure 3.1: Work Flow

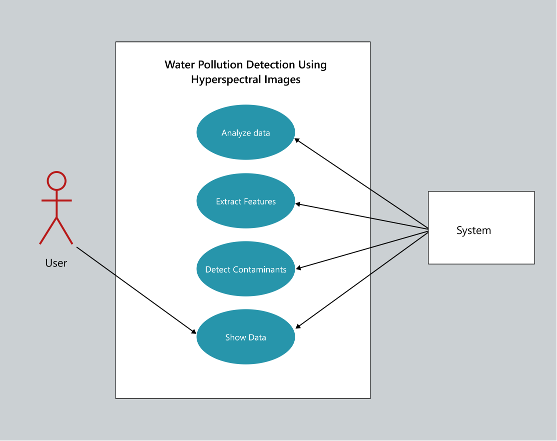


Figure 3.2: Use Case

3.3 Use Case Diagram

Diagram shows the whole dynamic aspect of the environment. It provides a graphical representation of what the system will do when undergoing specific or general requirements, shows the use case diagram of the system. See figure 3.2

3.4 Activity Diagram

A Use Case Diagram depicts the system's fundamental dynamic nature. It shows a graphical picture of the system's performance when subjected to certain particular or general constraints. The system's use case diagram is given below.

Chapter 4

Methodology

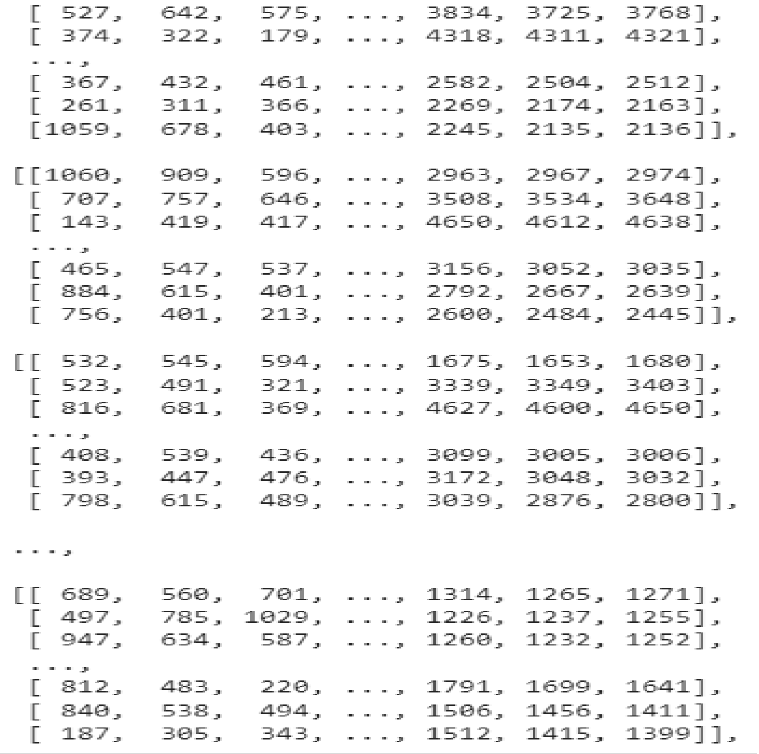
Data collection, analysis, feature extraction, and other procedures are part of a step-by-step approach to determining water pollution using hyperspectral pictures. A program is used to identify water contamination using hyperspectral photos. We will go through everything in depth further down.

4.1 Data Collection

There are no daily routine utilized data sources for Hyperspectral Images analysis, making it complex for newbies to start. However, we have completed this minor task. “The data was gathered from Pavia University. Over Pavia, Italy, a sensor known as the reflecting optics system imaging spectrometer (ROSIS-3). The data consists of 109 spectral bands. The size of HSI is 1096x715 pixels. The data contains a total of 9 classes. The total number of pixels are 783,640.

4.2 Data Preprocessing

Mat files are the most used format for Hyperspectral Image(HSI) data, which may be accessed using a variety of computer languages, including Python, as we did "one of the most significant preprocessing tasks is pixel extraction from "Hyperspectral Images".



```
  [ 527,  642,  575, ..., 3834, 3725, 3768],
  [ 374,  322,  179, ..., 4318, 4311, 4321],
  ...,
  [ 367,  432,  461, ..., 2582, 2504, 2512],
  [ 261,  311,  366, ..., 2269, 2174, 2163],
  [1059,  678,  403, ..., 2245, 2135, 2136]],

 [[1060,  909,  596, ..., 2963, 2967, 2974],
  [ 707,  757,  646, ..., 3508, 3534, 3648],
  [ 143,  419,  417, ..., 4650, 4612, 4638],
  ...,
  [ 465,  547,  537, ..., 3156, 3052, 3035],
  [ 884,  615,  401, ..., 2792, 2667, 2639],
  [ 756,  401,  213, ..., 2600, 2484, 2445]],

 [[ 532,  545,  594, ..., 1675, 1653, 1680],
  [ 523,  491,  321, ..., 3339, 3349, 3403],
  [ 816,  681,  369, ..., 4627, 4600, 4650],
  ...,
  [ 408,  539,  436, ..., 3099, 3005, 3006],
  [ 393,  447,  476, ..., 3172, 3048, 3032],
  [ 798,  615,  489, ..., 3039, 2876, 2800]],

 ...,

 [[ 689,  560,  701, ..., 1314, 1265, 1271],
  [ 497,  785, 1029, ..., 1226, 1237, 1255],
  [ 947,  634,  587, ..., 1260, 1232, 1252],
  ...,
  [ 812,  483,  220, ..., 1791, 1699, 1641],
  [ 840,  538,  494, ..., 1506, 1456, 1411],
  [ 187,  305,  343, ..., 1512, 1415, 1399]],
```

Figure 4.1: Date in dataset

This simplifies data management and machine learning techniques such as classification and clustering. The "Pavia University" Dataset is provided as an example. Certain pixels must be eliminated before the analysis since they contain no information. The geometric resolution is 13 meters. A few examples of bands from "Pavia University's" HSI are shown below".

4.2.1 Reading Dataset

This gives you “the data, the ground truth, or classes, as well as the data's and ground truth's sizes, which are 3D and 2D matrices, respectively”.

4.2.2 Extracting Pixels

“Pixels are individual elements in a Hyperspectral Image (HSI), which is a vector with a length equal to the number of bands in the HSI. We use code to extract pixels from HSI and save them to a CSV file for further use”. See figure **??**

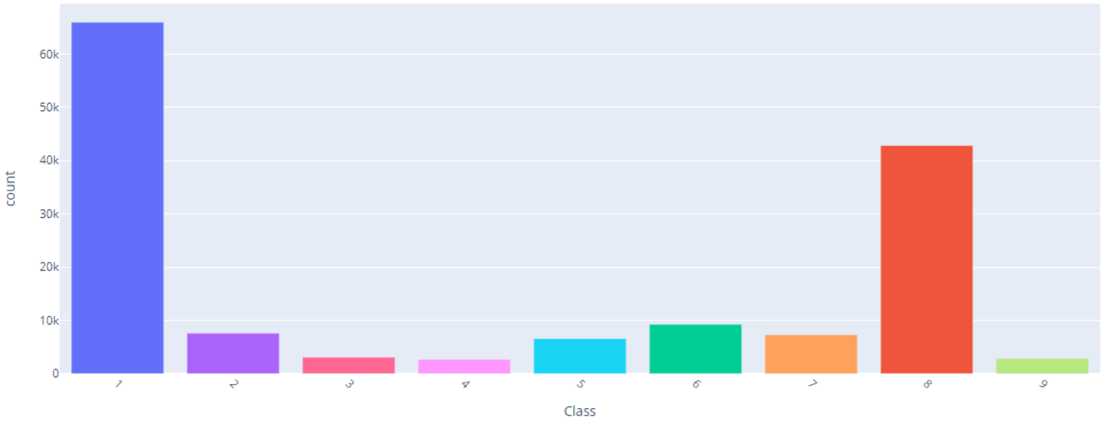


Figure 4.2: Bar Plot

4.3 Exploratory Data Analysis

It is challenging to analyze “the Pavia University dataset” since it has many dimensions. “As a consequence, using Principal Component Analysis (PCA), a popular and widely used dimensionality reduction technique, the data dimensions are reduced to 3D”(4). We had to limit the dataset’s dimensions to three.

4.3.1 Interactive Visualizations

The collection contains almost 783k patterns, it is not easy to see them all. As a result, visualizing the data is the most effective technique to see all data points, observations and pixels.

4.3.1.1 Bar Plot

Instead of a quantitative variable, it may be seen as a "histogram across a categorical variable". The graph below depicts the relationship between the HSI classifications. As we can see, "WATER" is the most popular class in the dataset. See figure 4.2

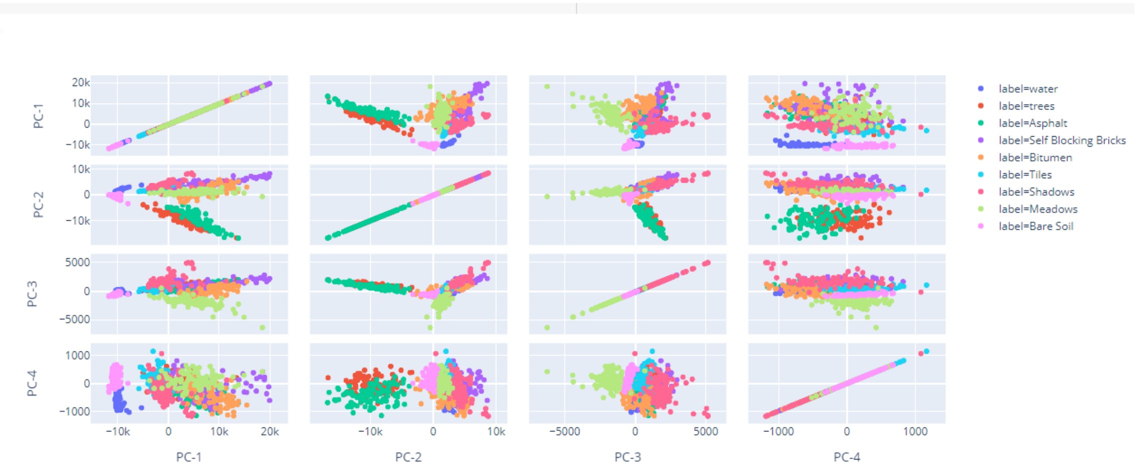


Figure 4.3: Pair Plot

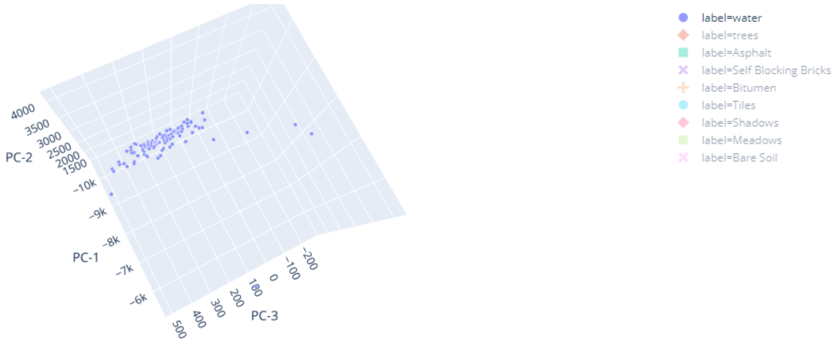


Figure 4.4: Scattered Plot

4.3.1.2 Pair plot

It is a straightforward method for visualizing the relationships between variables. It produces a relationship matrix for every variable in the dataset. The diagram depicts the relationship between the main components (PC1, PC2, and PC3). See figure 4.3

4.3.1.3 3D Scatter Plot

It displays the relationship between three variables by plotting data points on a 3D axis. “In the form of a 3D scatter plot, the image below depicts a relationship between main components (PC1, PC2, and PC3)”. See figure 4.4

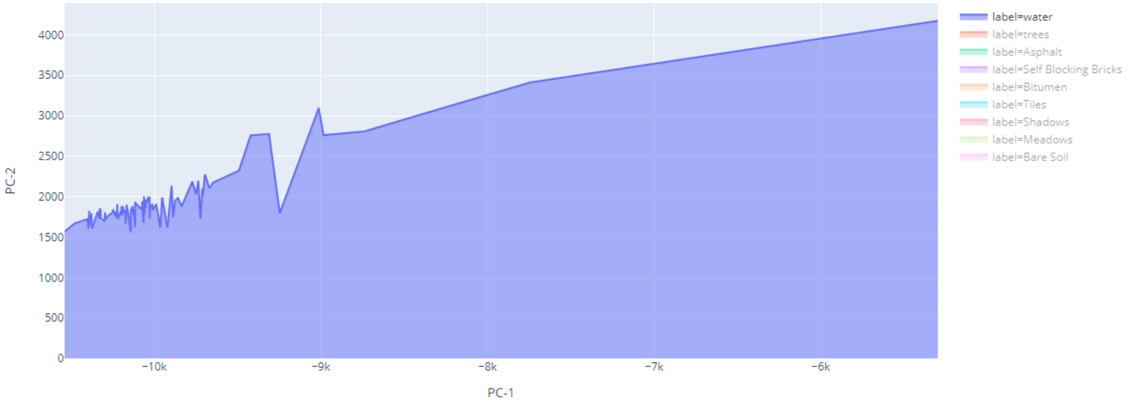


Figure 4.5: Area Plot

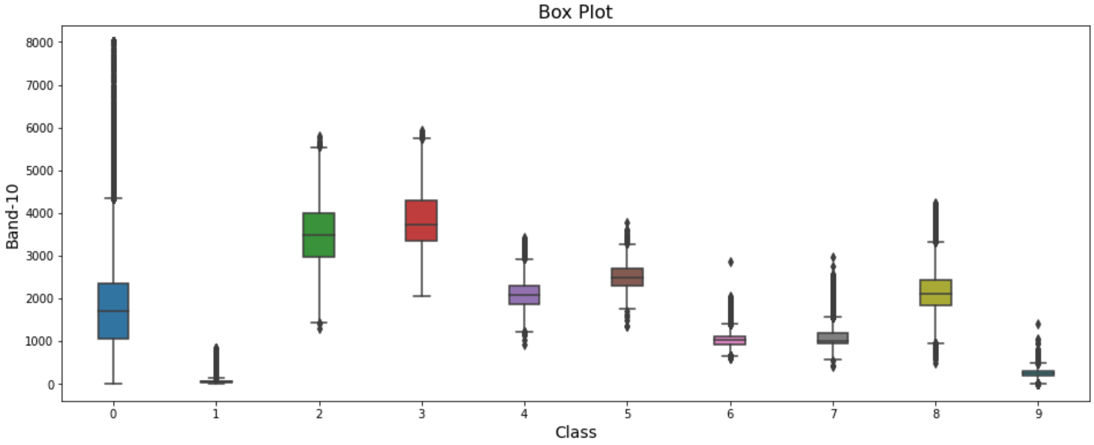


Figure 4.6: Box PLot

4.3.1.4 Area Plot

It shows the change of one variable in relation to another by connecting the data points with line segments. “The visualization with relation to the Principal Components (PC1, PC2, and PC3)”.See figure 4.5

4.3.2 Box Plot

It is a box and whisker plot representation. This plot usually represents 5 points "minimum", "first quartile", "median", "third quartile" and "maximum".“Visualization with relation to the Principal Components (PC1, PC2, and PC3)”. See figure 4.6

4.4 Dimensionality Reduction(DR)

“Dimensionality Reduction is a technique for reducing the number of dimensions in data, allowing classifiers to create complete models at a minimal computational cost. As a result, Dimensionality Reduction (DR) has gained prominence in order to increase the accuracy of pixel categorization in Hyperspectral Images (HSI)”.

Dimensionality Reduction has to types:

4.4.1 Feature Extraction

“Feature extraction is the process of identifying new features by selecting or combining existing features to produce a reduced feature space while still characterizing the data set appropriately and comprehensively.”

4.4.2 Feature Selection

The process of choosing dimensions of dataset features that help with machine learning tasks like classification and grouping. This may be achieved using a variety of approaches, including correlation analysis and univariate analysis.

“It is the process of selecting dataset properties that have a major impact on machine learning tasks such as classification, clustering, and so on. This may be accomplished using various methods, including correlation analysis and univariate analysis”.

4.5 Principal Component Analysis(PCA)

“PCA is an unsupervised linear dimensionality reduction approach that may be characterized as an eigen decomposition of the data’s covariance matrix. It improves interpretability by minimizing information loss by generating uncorrelated characteristics with maximum variance(4). The primary principle behind principal component analysis (PCA) is

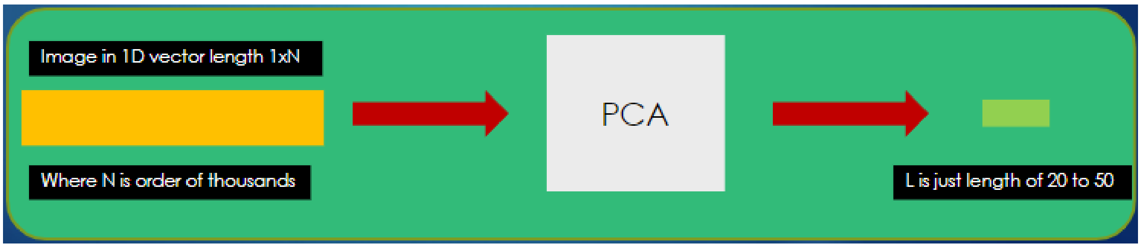


Image in 1D vector length 1xN
Where N is order of thousands
PCA
L is just length of 20 to 50

Figure 4.7: PCA

to minimize the dimensionality of a data set composed of many connected variables while keeping as much variance as feasible. This is accomplished by converting to a new collection of variables, the principal components (PCs), which are uncorrelated and organized in such a way that the first few maintain the majority of the variance included in all of the original variables”. See figure 4.7

4.5.1 Mathematics Behind PCA

“PCA is a type of unsupervised learning issue. The entire process of extracting principal components from a raw dataset may be broken down into six steps:

1. Take the whole dataset consisting of d+1 dimensions and ignore the labels such that our new dataset becomes d dimensional.
2. Compute the mean for every dimension of the whole dataset.
3. Compute the covariance matrix of the whole dataset.
4. Compute eigenvectors and the corresponding eigenvalues.
5. Sort the eigenvectors by decreasing eigenvalues and choose k eigenvectors with the largest eigenvalues to form a d × k dimensional matrix W.
6. Use this d × k eigenvector matrix to transform the samples onto the new subspace”.

4.5.2 How many dimensions should we reduce?

Using the "Explained Variance Ratio Graph", we can determine the appropriate number of dimensions to minimise.The significant components is represented on the x-axis, and

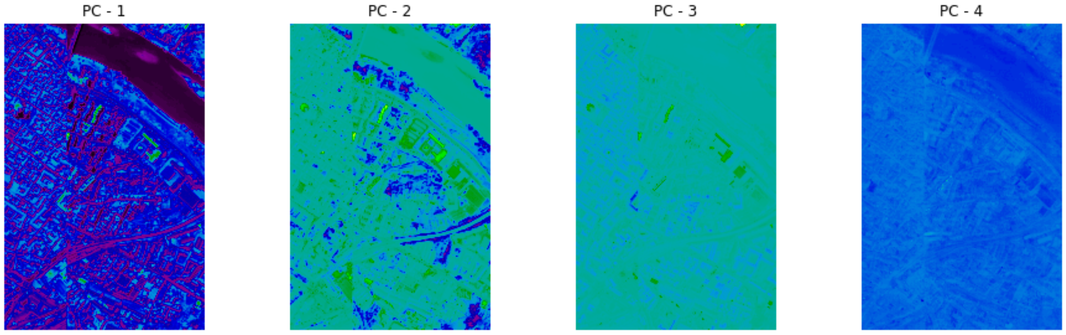


Figure 4.8: Extracted 4 band after DR

the cumulative explained variance is represented on the y-axis. We have the option of selecting the number of components with a cumulative explained variance more significant than 95 percent.

We now know how many dimensions of the HSI data we should remove. “Use Principal Component Analysis (PCA) to analyze the HSI data. Following the reduction of the data’s dimensions, class labels are added for future use”.

Let’s look at the four bands of the "Pavia University" HSI after dimensionality reduction with PCA. See figure 4.8

4.6 Classification Algorithms

Classification is a predictive modelling task that predicts a class label for given input data. The categorisation is as follows:

1. "Classification Predictive Modeling".
2. "Binary Classification".
3. "Multi-Class Classification".
4. "Multi-Label Classification".
5. "Imbalanced Classification".

We are now working on the "Multi-Class Classification challenge". “For the categorization of Hyperspectral Images (HSI), many classification techniques are utilised, including”:

1. "K-Nearest Neighbors"
2. "Support Vector Machine"
3. "Spectral Angle Mapper"
4. "Convolutional Neural Networks"
5. "Decision Trees"

We are going to use “the Support Vector Machine(SVM) to classify the Hyperspectral Image(HSI)”.

4.6.1 Support Vector Machine(SVM)

“SVM is a statistical learning theory-based supervised machine learning algorithm. To put it another way, SVM tries to locate a hyperplane in the multidimensional feature space to divide the two classes”. And this hyperplane is the optimal decision surface because it maximizes the margin, which is the distance between the hyperplane and two classes. “In general, the larger the margin, the better the classifier is. Solving an optimization issue for finding a hyperplane is analogous to obtaining an SVM model from a specified training set. SVM employs a minimum structural technique to avoid over-fitting problems in this optimization. With high-dimensional data and a limited training set, SVM performs effectively. According to Changjiang at el. (26) When it comes to HSI classification, several studies have revealed that SVM classifiers outperform other common classifiers, including decision tree classifiers, k-nearest neighbor classifiers, and neural networks. The kernel function of SVM, namely the radial basis function, is responsible for the majority of its power”.

Chapter 5

Results

5.1 VISUALIZING AND FEATURE EXTRACTION

5.1.1 BANDS

Imaging spectrometers are powerful sensor equipment that acquire spectral remote sensing data. The energy of reflected light is gathered in "bands" using imaging spectrometers. The electromagnetic spectrum is divided into bands, each of which represents a segment. See figure 5.1

5.1.2 GROUND TRUTH

The term "ground truth" refers to data gathered on the ground. Image data may be linked to real-world characteristics and materials using ground truth. A collection of measurements proven to be significantly more accurate than measures from the system you're analyzing is referred to as "ground truth." See figure 5.2

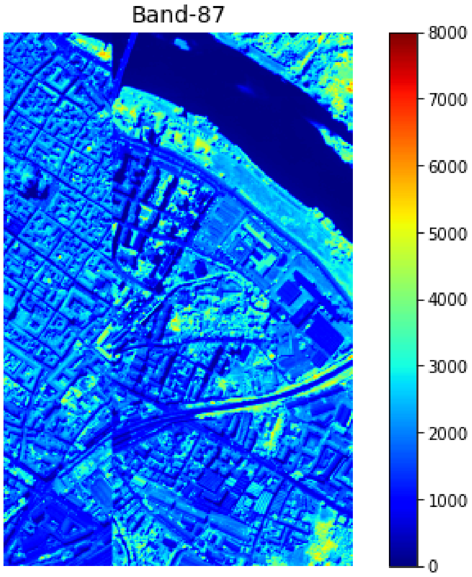


Figure 5.1: Band Extracted From Our Dataset

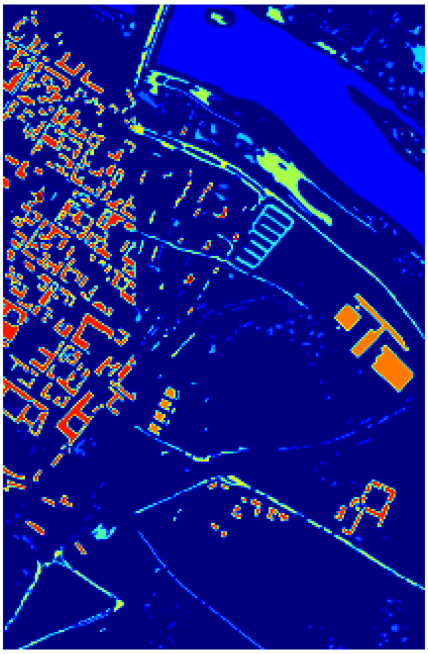


Figure 5.2: Ground Truth Extracted From Our Dataset

5.1.3 SPECTRAL SIGNATURE

"Spectral signatures", which are essentially plots of an object's spectral reflectance as a function of wavelength, are useful for image categorization since they include both "qualitative and quantitative information". See figure 5.3

5.1.4 CUMULATIVE VARIANCE

The amount of variance explained by each type of model is plotted compared to the total number of components. See figure 5.4

5.1.5 BANDS AFTER PCA

The band changed after we applied the "PCA" and we get the bands. See figure 5.5

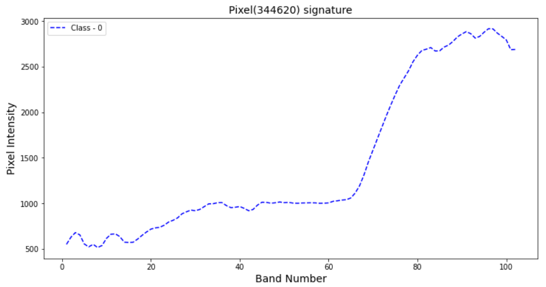


Figure 5.3: Spectral Signature of Band

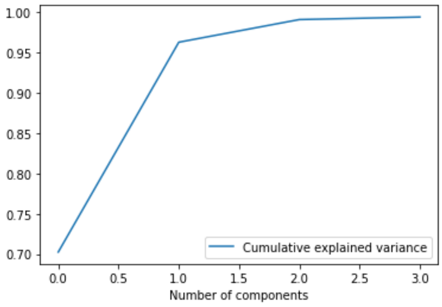


Figure 5.4: Cumulative Variance Between Band

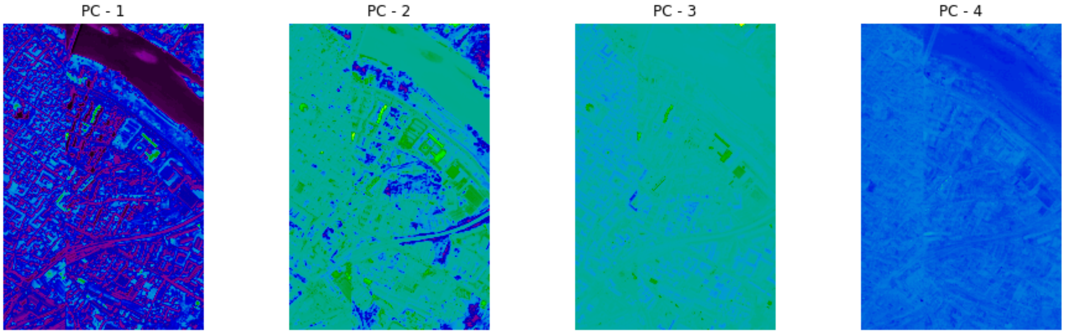


Figure 5.5: Extracted Band After PCA

5.1.6 CONFUSION MATRIX

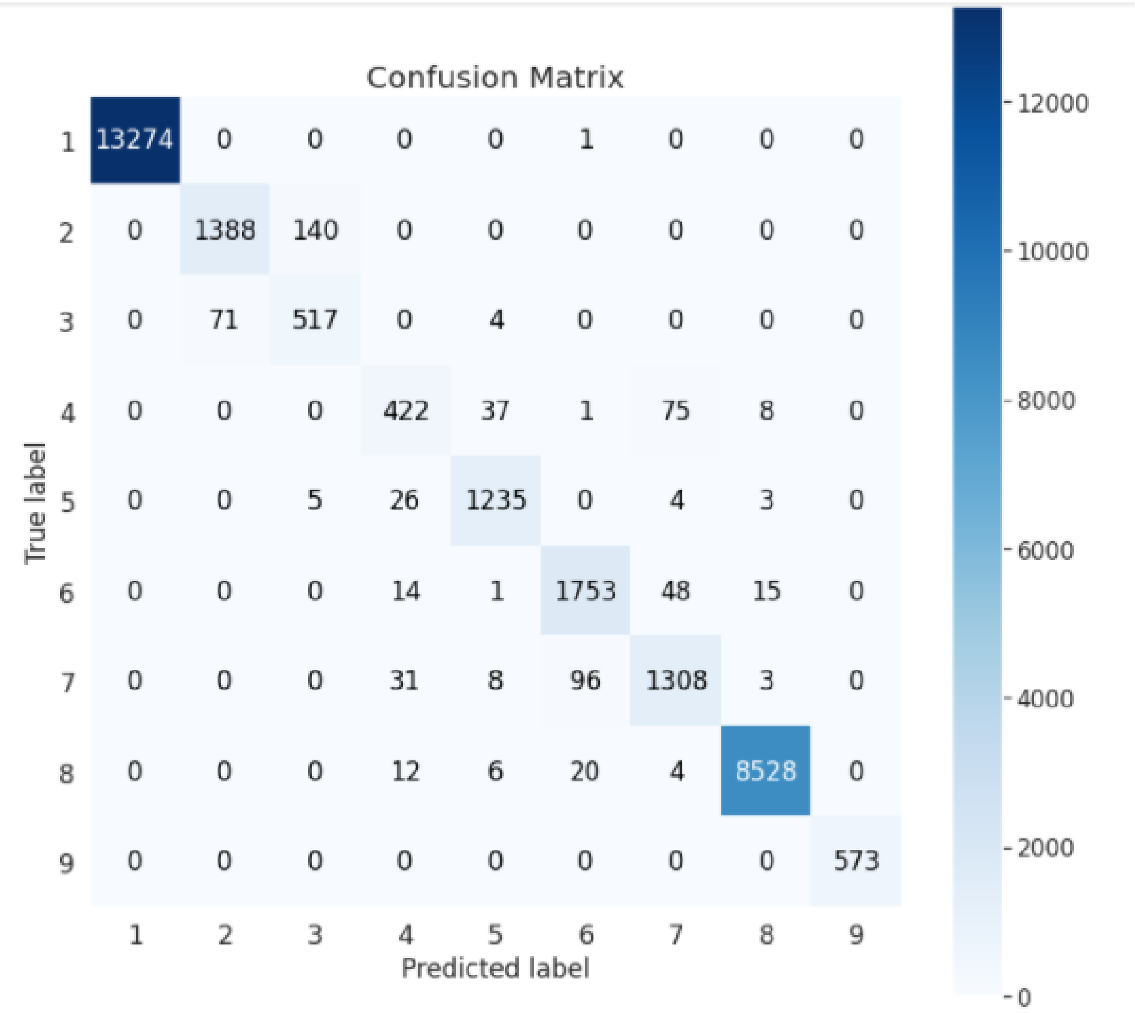
The "quantitative technique" of characterising image categorization accuracy is usually a "confusion matrix." It's a table that shows how the "classification result and a reference image" correlate. See figure 5.6

5.1.7 CLASSIFICATION REPORT

The Classification Report created, which includes “Class-wise Accuracy, Accuracy Precision, Recall, F1 Score, and Support, is presented in figure 5.7

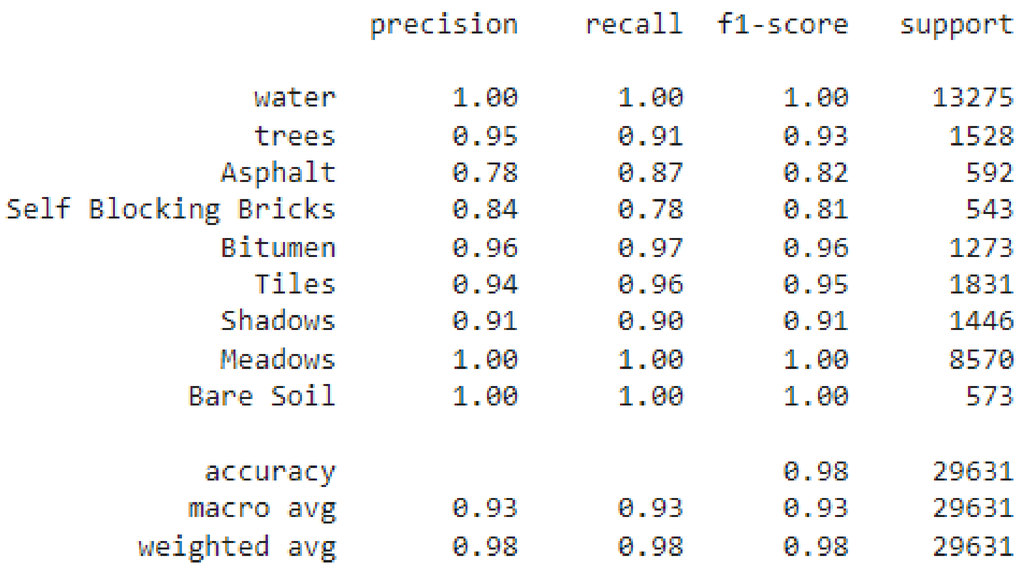
5.1.8 MODEL ACCURACY

The model's accuracy is determined by how often it properly classifies an image. The accuracy for our model was around 97 percent. Accuracy: 0.9786372380277412



	1	2	3	4	5	6	7	8	9
1	13274	0	0	0	0	1	0	0	0
2	0	1388	140	0	0	0	0	0	0
3	0	71	517	0	4	0	0	0	0
4	0	0	0	422	37	1	75	8	0
5	0	0	5	26	1235	0	4	3	0
6	0	0	0	14	1	1753	48	15	0
7	0	0	0	31	8	96	1308	3	0
8	0	0	0	12	6	20	4	8528	0
9	0	0	0	0	0	0	0	0	573

Figure 5.6: Performance Of Model



```
                    precision    recall  f1-score   support

             water       1.00      1.00      1.00     13275
             trees       0.95      0.91      0.93      1528
           Asphalt       0.78      0.87      0.82       592
Self Blocking Bricks       0.84      0.78      0.81       543
           Bitumen       0.96      0.97      0.96      1273
             Tiles       0.94      0.96      0.95      1831
           Shadows       0.91      0.90      0.91      1446
           Meadows       1.00      1.00      1.00      8570
         Bare Soil       1.00      1.00      1.00       573

          accuracy                           0.98     29631
         macro avg       0.93      0.93      0.93     29631
      weighted avg       0.98      0.98      0.98     29631
```

Figure 5.7: Quality Of Prediction

5.1.9 FINAL BAND

After the Final "PCA" technique we got the band in which we got the water samples, See figure 5.8

5.2 EXPERIMENTAL RESULTS

"Phosphorus" , "nitrogen", "COD", "BOD", "turbidity", and "Chl-a" content levels in this research vary from 0.09 mg/L to 0.52 mg/L, 0.09 mg/L to 5.37 mg/L, 5.0 mg/L to 58.0 mg/L, 1.0 mg/L to 13.9 mg/L, 10 NTU to 97 NTU, and 3 g/L to 238 g/L, respectively because the water was obtained in fish breeding pools, the "turbidity", "Chl-a", "BOD", "COD", and "nitrogen" are the greatest. The high levels of BOD, "COD", and "nitrogen" are caused by organic materials. "Turbidity is heavily concentrated in pools due to a lack of water exchange, causing turbidity to rise because of the presence of water exchange and little living wastes, other routes have a low concentration of water quality indicators.".(3)

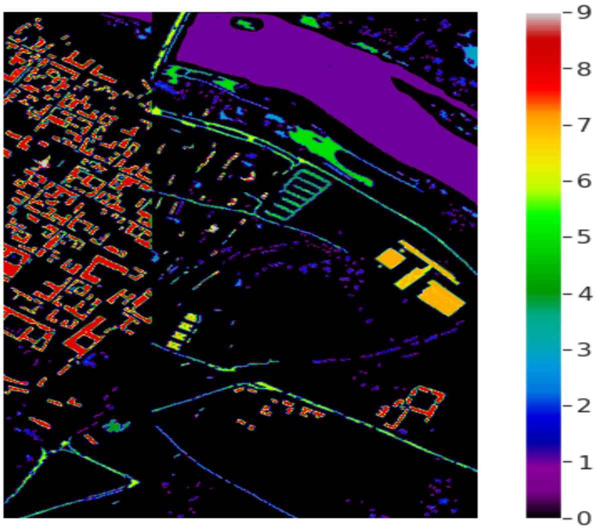


Figure 5.8: Final Band Extracted After PCA

Chapter 6

Discussion

The paper's main goal is to look at the possibility of estimating five distinct water quality parameters using solely measured, and hence sparse, input data. Previous research have devoted little emphasis to the use of machine learning in estimating water quality indicators using hyperspectral data. Regressions may be performed using machine learning without knowing anything about the water body or the water quality metrics being studied. Furthermore, unlike band-ratio approaches, this technique is entirely data-driven, requiring no new features to be developed based on domain knowledge.

Chapter 7

Conclusion and Future Work

Estimating the "chlorophyll-a", "Nitrogen" and other seawater parameters from remote sensing images is still an open and hot research topic. Many satellite images are used to estimate "chlorophyll-a" and "Nitrogen" from coarse low-resolution images that are not even considered for the study of complicated nonlinear data like coastal waters. Other satellite image types include medium and high resolution "Multispectral" and "Hyperspectral images", with the latter being more accurate in estimation than the former due to acceptable and contiguous spectral resolution of less than "10 nm" between bands, compared to more than "80 nm" for medium resolution "Multispectral images".(27)
Hyperspectral systems, which create more comprehensive spectral data, have made it possible to combine several hundred spectral bands in a single collection. "Multispectral Imagery" has been the only data source in air and water observational remote sensing from aerial or satellite operations since the 1960s, until improvements in hyperspectral remote sensing. (9) However, multispectral remote sensing data from the visible close and ultraviolet infrared areas of the visible spectrum were only gathered in 3 to 6 spectral bands in a particular sample, making it difficult to examine water quality from this data source. The hyperspectral remote sensing data processing employing field spectrometer data and remote sensing of water quality were discussed in this chapter.
In this project, we gather the dataset from Pavia University. We had different classes, around 9, which were useless for our work in that dataset. We have reduce the number of the band for our ease. The technique we used to reduce our data's dimensions was Princi-

pal Component Analysis (PCA). We had reduced our dimensions from 109 to 4 to remove the unwanted signals from the images. Now then, we wanted to detect the water from the given hyperspectral images, so we used Support Vector Machine to train the model to find out only the water to carry on our project to find our impurities from it. After that, we applied the ANN machine learning model to find different impurities from the water like "Chlorophyll-a" , "Nitrogen", and many more. This project aimed to detect impurities using hyperspectral images from any dataset. The results that we have achieved can be improved further by using the latest and high-quality images.

After working on this project, we concluded that wavelengths correlate with the human activity of different materials. These wavelengths carry some anonymous information that we still do not know of. We have worked on detecting impurities and predicting some impurities, and we believe that more information regarding impurities can be extracted from these wavelengths.

This project can be further advanced to detect impurities in water and converted to detect many other materials like detection mines installed underground. A slight change can bring a great novelty in this project for future work and open a vast field for many upcoming students like us.

Bibliography

[1] C. Cheng, Y. Wei, X. Sun, and Y. Zhou, “Estimation of chlorophyll-a concentration in turbid lake using spectral smoothing and derivative analysis,” *International Journal of Environmental Research and Public Health*, vol. 10, no. 7, p. 2979–2994, Jul 2013. [Online]. Available: http://dx.doi.org/10.3390/ijerph10072979

[2] M. Awad, “Sea water chlorophyll-a estimation using hyperspectral images and supervised artificial neural network,” *Ecological informatics*, vol. 24, pp. 60–68, 2014.

[3] Y. Zhang, L. Wu, H. Ren, Y. Liu, Y. Zheng, Y. Liu, and J. Dong, “Mapping water quality parameters in urban rivers from hyperspectral images using a new self-adapting selection of multiple artificial neural networks,” *Remote Sensing*, vol. 12, no. 2, p. 336, 2020.

[4] S. Kakarla, “Hyperspectral image analysis-getting started,” *Mathematical Modeling and Analysis*, 2021.

[5] R. Boellaard, “Es 04-01—image processing and analysis using artificial intelligence (cnns).” *Radiology*, vol. 291, pp. 53–59, 2019.

[6] N. G. Rostom, A. A. Shalaby, Y. M. Issa, and A. A. Afifi, “Evaluation of mariut lake water quality using hyperspectral remote sensing and laboratory works,” *The Egyptian Journal of Remote Sensing and Space Science*, vol. 20, pp. S39–S48, 2017.

[7] M. Mbuh, “Use of hyperspectral remote sensing to estimate water quality,” in *Processing and Analysis of Hyperspectral Data*. IntechOpen, 2019.

[8] W. K. Dodds and V. H. Smith, “Nitrogen, phosphorus, and eutrophication in streams,” *Inland Waters*, vol. 6, no. 2, pp. 155–164, 2016.

[9] M. W. Matthews, S. Bernard, and K. Winter, "Remote sensing of cyanobacteria-dominant algal blooms and water quality parameters in zeekoevlei, a small hypertrophic lake, using meris," *Remote sensing of environment*, vol. 114, no. 9, pp. 2070–2087, 2010.

[10] S. Keller, P. M. Maier, F. M. Riese, S. Norra, A. Holbach, N. Börsig, A. Wilhelms, C. Moldaenke, A. Zaake, and S. Hinz, "Hyperspectral data and machine learning for estimating cdom, chlorophyll a, diatoms, green algae and turbidity," *International journal of environmental research and public health*, vol. 15, no. 9, p. 1881, 2018.

[11] J. Tan, K. A. Cherkauer, and I. Chaubey, "Using hyperspectral data to quantify water-quality parameters in the wabash river and its tributaries, indiana," *International Journal of Remote Sensing*, vol. 36, no. 21, pp. 5466–5484, 2015.

[12] X. Wang and W. Yang, "Water quality monitoring and evaluation using remote sensing techniques in china: a systematic review," *Ecosystem Health and Sustainability*, vol. 5, no. 1, pp. 47–56, 2019.

[13] R. Bhateria and D. Jain, "Water quality assessment of lake water: a review," *Sustainable Water Resources Management*, vol. 2, no. 2, pp. 161–173, 2016.

[14] N. Usali and M. H. Ismail, "Use of remote sensing and gis in monitoring water quality," *Journal of sustainable development*, vol. 3, no. 3, p. 228, 2010.

[15] B. Mcilwaine, M. R. Casado, and P. Leinster, "Using 1st derivative reflectance signatures within a remote sensing framework to identify macroalgae in marine environments," *Remote Sensing*, vol. 11, no. 6, p. 704, 2019.

[16] P. M. Maier and S. Keller, "Machine learning regression on hyperspectral data to estimate multiple water parameters," in *2018 9th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*. IEEE, 2018, pp. 1–5.

[17] J. F. Schalles, A. A. Gitelson, Y. Z. Yacobi, and A. E. Kroenke, "Estimation of chlorophyll a from time series measurements of high spectral resolution reflectance in an eutrophic lake," *Journal of Phycology*, vol. 34, no. 2, pp. 383–390, 1998.

[18] D. C. Rundquist, L. Han, J. F. Schalles, and J. S. Peake, "Remote measurement of algal chlorophyll in surface waters: the case for the first derivative of reflectance near 690 nm," *Photogrammetric Engineering and Remote Sensing*, vol. 62, no. 2, pp. 195–200, 1996.

[19] D. J. Suggett, O. Prášil, and M. A. Borowitzka, *Chlorophyll a fluorescence in aquatic sciences: methods and applications*. Springer, 2010, vol. 4.

[20] P. Brezonik, K. D. Menken, and M. Bauer, "Landsat-based remote sensing of lake water quality characteristics, including chlorophyll and colored dissolved organic matter (cdom)," *Lake and Reservoir Management*, vol. 21, no. 4, pp. 373–382, 2005.

[21] A. G. Dekker, "Detection of optical water quality parameters for eutrophic waters by high resolution remote sensing," 1993.

[22] F. L. Hellweger, W. Miller, and K. S. Oshodi, "Mapping turbidity in the charles river, boston using a high-resolution satellite," *Environmental monitoring and assessment*, vol. 132, no. 1, pp. 311–320, 2007.

[23] A. A. Gitelson, Y. Z. Yacobi, A. Karnieli, and N. Kress, "Remote estimation of chlorophyll concentration in polluted marine waters in haifa bay, southeastern mediterranean," in *Air Toxics and Water Monitoring*, vol. 2503. International Society for Optics and Photonics, 1995, pp. 44–54.

[24] B. Hakansson and M. Moberg, "Cover. the algal bloom in the baltic during july and august 1991, as observed from the noaa weather satellites," *TitleREMOTE SENSING*, vol. 15, no. 5, pp. 963–965, 1994.

[25] L. Han and K. J. Jordan, "Estimating and mapping chlorophyll-a concentration in pensacola bay, florida using landsat etm+ data," *International Journal of Remote Sensing*, vol. 26, no. 23, pp. 5245–5254, 2005.

[26] C. Liu, F. Zhang, X. Ge, X. Zhang, Y. Qi *et al.*, "Measurement of total nitrogen concentration in surface water using hyperspectral band observation method," *Water*, vol. 12, no. 7, p. 1842, 2020.

[27] A. I. Flores-Anderson, R. Griffin, M. Dix, C. S. Romero-Oliva, G. Ochaeta, J. Skinner-Alvarado, M. V. Ramirez Moran, B. Hernandez, E. Cherrington, B. Page *et al.*, “Hyperspectral satellite remote sensing of water quality in lake atitlán, guatemala,” *Frontiers in Environmental Science*, vol. 8, p. 7, 2020.