Study on Machine Learning Algorithms and Statistical Analysis for Classification of Hydrothermal Seafloor Rocks Measured Underwater Using Laser-Induced Breakdown Spectroscopy

DISSERTATION

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY in ENGINEERING

YELAMELI, Mallikarjun Rajendrakumar

Supervisor: Prof. Kazuo Ishii

Department of Human Intelligence Systems

Graduate School of Life Science and System Engineering

Kyushu Institute of Technology

Japan

March 2019

Abstract

The aim is to study the use of machine learning algorithms for the classification hydrothermal seafloor rocks measured underwater using Laser-Induced Breakdown Spectroscopy. The rocks were classified concerning their labels assigned to each rock, and geological groups formed ternary diagram with the relative ratio of Cu-Pb-Zn. In this research the target rocks are obtained from deep-ocean in Okinawa Japan. These were hydrothermal deposit sea-floor rocks.

Further, these rocks were brought into the laboratory and broken into pieces and made the pellets. The experimental setup which resembles the ocean, then to test using ChemiCam device which is a LIBS device which is specifically designed for the chemical elemental analysis in the deep ocean is used to fire the laser beams on rocks. The proposed methods for classification of rocks with respect to their labels and for geological group are evaluated using with and without linear detrend along with the principal component analysis (PCA) as a pre-processing step which significantly reduces the dimensionality of the data, with classification algorithms such as the support vector machine (SVM), k-nearest neighbor search (KNN) and artificial neural network (ANN) methods. The performance of the classification algorithms depends on the size of the dataset, to prove this the dataset has been divided into two sets of 100 laser shots of each rock and 300 laser shots of each rock. Additionally, removing the noise from the spectra such as linear trend using linear detrend operation from the data enhances the performance of the classification in terms of sensitivity. The best classification performance concerning the rock label concerning

sensitivity is obtained using an SVM linear kernel algorithm with 95%. The best classification performance concerning the geological group is obtained using the SVM method with 98% accuracy. The one-sided Wilcoxon signed rank test is applied to the classification results in the rock label and group cases, and the results indicate that the SVM algorithm has statistical significance over the other algorithms while classifying the rock labels and rock group.

Keywords: Seafloor deposits, Classification, PCA, SVM, KNN, ANN, LIBS

List of Tables

List of Figures

Table of Contents

1 Introduction

1.1 Aim and highlights of this study

The manufacture of hi-tech applications such as a smartphone, hybrid cars, photovoltaic installations and other goods and types of machinery requires abundant industrial minerals and metals. Today the demand for industrial minerals and metals are multiplying due to rapid growth and growing need of BRIC (Brazil, Russia, India and China) countries. Today most of the demands of metals and industrial minerals are derived from onshore mining but to satisfy the growing need; scientist predicts fear of a shortage shortly. To overcome this problem and to meet the demand in future, the scientists started to look for the alternative option, and one of the practical and feasible options is deep ocean mining, though it is still too expensive. The ocean mining sites are around hydrothermal vents at about 1000 – 3000 metres below the ocean surface. The vents create sulphide deposits, and it contains various minerals such as copper, manganese, cobalt and zinc. The survey of the location using before mining is one of the critical steps, it helps to understand the minerals and chemical present in surrounding that area beside this it helps to understand the deep sea geochemical processes. Such types of surveys are carried out using a remotely operated vehicle (ROV) with modern chemical sensors.

NASA has developed the Chemcam device using LIBS as a major device for in situ chemical analysis on Mars[1]. On similar grounds, of Chemicam device, the Chemicam is successfully developed for analysis in the ocean, and it is a modern oceanographic

sensor[2]. The LIBS works on the basic principle of atomic emission spectroscopy. The high energy and ultra-short beam is focused on breaking the molecular bonds of the compound and forms a plasma; this emits specific wavelength. LIBS analyses solid, liquid and gas matter regardless of its physical state.

The hydrothermal rocks were obtained from the deep ocean using ROV from the site Okinawa Japan. These rocks were labelled concerning its geological location (latitude and longitude information) from where obtained. These rocks are further investigated in the lab. These rocks are broken into pieces and then crushed till the fine powder is obtained. Further, these rocks were pressed at pressure to make pellets. An artificial experimental setup was made in the lab which resembles the ocean. The pellet was dip into a container filled with ocean water. The LIBS device is also dipped into the ocean water container. The laser beam originated in the container filled with ocean water, strikes on pellet which is also in the container filled with ocean water; the plasma gets generated in the water, that plasma was recorded using ICCD camera. This experiment was repeated for 30 rocks.

In this thesis, the method has been proposed to classify the rocks not only concerning its label but also concerning geological group it belongs[3]. The machine learning algorithms have been investigated those are SVM, KNN and ANN. The performance of these algorithms have been compared using one-sided Wilcoxen signed rank test, and the best algorithm has been proposed to classify the rocks not only concerning its label but also concerning the geological group.

1.2 Energy requirement

Energy is the critical requirement for the growth of any society or country. One of the prominent energy resources is natural gases, oil, metals and minerals. The state which has more reservation of these natural resources has great ability to exert influence or project power on a global scale. No society or state can flourish without enough energy resources. The applications of these energy resources are ranges in every aspect of our life right from fuel required for cars to the airplane to travel, in the manufacturing industries such as a smartphone, computer hardware industries, cars and electronic devices industries which require industrial metals. Apart from these, the energy needed to operate any devices right from a smartphone, iPad, computer, laptop and to generate the electricity from coal, trains and aeroplanes to travel. It is almost impossible to imagine life without the need for energy. The demand for these natural gases, oils, minerals and metals has been increasing day by day, and the available resources are shrinking day by day[4]. Till today the energy demands were met by onshore mining, but it is time to look for alternative resources so that that requirement will be fulfilled in the near future. One of the prominent alternative to extracting these energy resources is ocean mining also known as offshore mining[5].

It takes millions of years to form the gases and oils in the ocean. It is formed by flushed particles from the land, buried and compressed in the layers of several kilometres, the earth atmosphere, pressure and temperature conditions; bacteria convert the biomass into precursor substance to hydrocarbons[4]. These hydrocarbons spread over specific layers to form the rock and sediments[4]. Today's natural resources are between 15 to 600 million

years old. Oil and gas are usually found where vast layers of sediment cover the ocean floor[4].

As ocean covers almost over 70% of the earth. Usually, to find out the location of these natural gases, seismic equipment's are used, which transmits the sound waves which get reflected from the layers of rock and sediments. From the sounds, geologists can estimate whether the layers could contain oil or natural gas[5].

1.3 Marine Minerals

The ocean is a treasure of valuable energy resources. The oil and gas have been extracted from the ocean for many years. Now, the primary focus is on manganese noodles which are usually located at a depth of below 4000m, gas hydrate located around 350 to 5000 meters and cobalt crust which ranges between 1000 and 3000 meters[5].

The manganese noodles are composed of iron, silicates and hydroxides. These elements are of economic interest. The manages noodles are usually found in the west coast of Mexico, Peru, Basin and the Indian ocean. The cobalt crusts form at the depths of 1000 to 3000 meters on the flanks of volcanoes and therefore usually occur in submarine volcanoes. These submarine volcanoes consist of copper, nickel and platinum. Massive sulphides are sulphur-rich ore that originates at the "black smokers" are also of great interests[5].

1.4 Hydrothermal Vents

Hydrothermal vents were discovered in the Pacific in 1979; they are metal-bearing sulphur compounds[5]. Besides hydrothermal vents, there are manganese nodules and cobalt which

consists of metal-bearing mineral resources. Massive sulphides originate at hot vents in the ocean due to sulphide-enriched water flows out of the seabed[5]. One of the active hydrothermal vent images is shown in Fig.1.1. The hydrothermal vents are found along plate boundaries and active volcano in the ocean[5]. Table 1.1 shows the details of hydrothermal vents in the ocean.

1.4.1 Basics of Hydrothermal Vents

It forms due to heat exchange between crustal rocks and the ocean. Seawater penetrates the seafloor for several thousand meters. At these depths, the seawater is heated to due to temperatures of around 400 degrees Celsius by magma. It undergoes many chemical changes which bring nutrients to the sea-floor. These nutrients are essential because bacteria need them to perform chemosynthesis[6].

Chemosynthesis is a process somewhat similar to photosynthesis but in the photo without the use of sunlight, after the water seeps through the crust oxygen and potassium are removed from the seawater as the water sinks deeper, sodium, potassium and calcium into the water from the surrounding crust. Even more in-depth in the crust, water reaches its highest temperature from being heated by the hot magma and other element dissolve in the water such as zinc, iron, hydrogen sulphide and hydrogen[6] as shown in Fig. 1.2[7].

Next, the hot fluid carrying the metals starts to rise towards the crust. The hot hydrothermal fluids mix with the cold seawater as they are emitted from the hydrothermal vents and form metal sulphide materials as shown in fig. 1.2[7].

These warm fluids are rich in chemicals so chemosynthetic bacteria can use them for chemosynthesis they are the primary producers in this system, and they utilise sulphur, hydrogen, methane and other compounds released by the reactions between sea-water and magma.

The organisms that thrive near these vents are called extremophiles. These extremophiles can survive extreme temperatures, pressures, salinities are hostile to normal life.

1.4.2 Global distribution of hydrothermal vents.

Hydrothermal vents have been found most of the places in the world map. They get formed between the range of 1000 to the 4000m range. As shown in Fig. 1.3[8] there are many locations where one can find the hydrothermal vents. There are many confirmed hydrothermal vents around east pacific rise. According to a geologist, there are four common areas of origin for hydrothermal vents[5].

- At Mid Ocean Ridges: These are mountain ranges in the ocean that circle the earth like the seam on a baseball. At these locations, the oceanic plates drift apart, and this drift produces the fractures through which water sinks.
- At Island-Arc Volcanoes: When one oceanic plate is forced beneath another one under the sea, these volcanoes get formed.
- Volcanoes behind island arcs (black arc basins): These are formed when one plate submerges beneath another, tension is produced in the overlying plate.

- At intraplate volcanoes: These volcanoes form in the interiors. The Hawaiian island group is an example.

Deep sea mining is an extremely challenging task and since the operation is being conducted in deep ocean surface where the environmental conditions are at extreme and using the remote sensing tool is one of the options. The laser-induced breakdown spectroscopy (LIBS) is one of the remote sensing devices which has been developed by the University of Tokyo[2].

1.5 In-situ chemical sensors

Since due to extreme environmental situation in the deep sea, the in-situ chemical sensors are used. There are various advantages of using an in-situ chemical sensor such as it can cover the larger area with a maximum number of sampling, with high spatial resolution. Some of the in-situ chemical sensors are sensors with electrodes such as pH sensor [9]. There are other sensors as well such as optical sensors. Optics widely used in-situ for monitoring physical, chemical and biological parameters in the deep sea [10]. Mass spectroscopic sensors have developed for multi-elements in gas sea-water[11], [12]. Laser Raman (LR) spectroscopy has been applied to deep-sea surveys; it is a technique based on inelastic scattering of light of vibration of the molecule [13]. The LR spectroscopy has been successfully deployed in the deep ocean for the measurement of rocks and vent fluids at hydrothermal vents [14].

1.6 Laser-induced breakdown spectroscopy

To understand the Chemicam device, it is essential to understand the laser-induced breakdown spectroscopy(LIBS). LIBS is a chemical sensing device. Though there are several sensors, for example, Raman sensor for in-situ chemical analysis of the chemicals, gases in liquids but in-situ analysis of solids remains limited[15]. Laser-induced breakdown spectroscopy is one of such sensors. It is a tool to provide a fast and in situ chemical analysis that determines the elemental composition of the target. The advantage of using laser-induced breakdown spectroscopy that it does not need sample preparation[2]. It is extremely fast in measurement time. It usually takes a few seconds for analysis of a single spot. It covers a wide range of elements such as H, Be, Li, C, N, O, Na and Mg[16]. It is a form of atomic emission spectroscopy that analyses the light emitted from atom and ions of material from the plasma emitted by focussing the laser beam on the target. In the first stage, the plasma is highly ionised and then electron-ion recombination[17]. The light emitted from the spectrometer is recorded using the intensified charged-coupled device (ICCD) for elemental analysis. The plasma produced for each atom and ion varies. The significant advantages of LIBS are that it requires no sample preparation; the results are available in real time.

1.6.1 LIBS Applications

LIBS has been used for many in-situ elemental analysis which includes a solid, a liquid or a gas, and has wide applications in the medical science[18], geomaterials[19],

explosives[20], recycling[21], forensics[22], agriculture[23] and nuclear sectors[24]. Considering overall accuracy, it can be said that the suitability of LIBS for in-situ analysis is comparable to XRF [25]. LIBS has a great advantage in application to aquatic environments since XRF suffers from great attenuation of X-rays in water. The main reason for this LIBS popularity is its detection ability[26]. The spatial confinement and fast discharge enhance the signal multiple times[26]. So it has become famous and promising analysis method in the steel industry[21]. The LIBS application has been demonstrated in the food industry for the inspection of the wheat, barley etc. LIBS has shown element tracing at ppm level in the starch-based food samples with an acceptable precision[27].

1.6.2 The working principle of LIBS

The Fig. 1.4 shows the general setup for laser-induced breakdown spectroscopy. A mirror transfers the pulsed laser radiation to a focusing lens. The sample to be analysed is placed in a rotating sample holder. The laser beam is being adjusted in such a way using a mirror that if directs on sample perpendicularly to the sample surface. The focused laser beam generates the plasma at the sample surface. The emitted plasma is collected using collecting lens as shown in Fig. 1.4. The control unit generates the laser. The monochromator collects the signal. To improve the signal to noise ratio, the plasma radiation is recorded only during the lifetime of the plasma.

1.6.3 The working principle of plasma

In principle, whenever the laser strikes on the object, the electrons get excited, and it transits to a higher level as shown in Fig.1.5. However, the excited electron will be unstable and cannot remain at the higher state and returns to its original state while transiting to its original position it emits the photon which will be recorded. If we plot the intensity of photon along with the wavelength, we will get the characteristics of the material.

1.7 ChemiCam Device

ChemiCam is a laser-induced breakdown spectroscopy device (LIBS). LIBS is a rapid chemical analysis device. It uses a laser pulse and creates micro plasma on the target sample. The laser is a Q-switched Nd: YAG DPSS laser with a wavelength of 1064nm. It has maximum pulse energy is 20mJ, with the frequency of 2Hz. It has a pulse duration of 150ns. In the next generation device, the pulse duration can be raised up to 250ns. The laser system is consisting of two photodiodes that monitor the laser-pulse characteristics. The Fig. 1.6[2] shows the ChemiCam device.

The plasma generated underwater exist for a concise time, not more than 2.5μs. The intensity of the plasma generated in the air is higher than the intensity of the plasma generated in the water. The custom built Czerny-Turner spectrometer is used in the ChemiCam device. It has light throughput of f/4.5. The ICCD model with high sensitivity in the spectral range was used. The range and resolution of the spectrometer were chosen based on preliminary experiments using rocks.

The optical system of ChemiCam has a high efficiency of light collection by separating fibres for laser delivery and plasma observation. One 600-µm-core fibre is used for delivery of laser light to the target's surface and multiple bundled 100--µm-core fibres for observation of light emitted from the plasma. The CPU controls and communicates with the laser, ICCD camera and other components and stores data. Fig 1.7 shows the long pulse laser beam. The deep sea LIBS sensor has been developed over the years by improving the performance and adding new features. It was all started in the year 2012; the first model was known as I-SEA (IN-situ Seafloor element analyser). The length was 1.5m diameter 0.3m and weight of 110kg in the air. It has two lasers, a double pulse technique, a spectrograph, ICCD for a detector and CPU[16]. The next generation was of the sensor is Chemicam developed in 2013. It is slightly smaller than the I-SEA, with a length of 1.3m and a diameter of 0.3m. The weight is more than the I-SEA around 150 Kg in the air[16]t. In this thesis 4th generation of the LIBS, the sensor device has been used as shown in fig. 1.6 and Fig. 1.7[2] The detail specification is listed in Table 1.2 [15]

1.8 Literature Review

One of the purposes of usage of machine learning algorithms in LIBS is classified as the material on an elemental basis. Obtaining the desired information is one of the significant challenges. This can be achieved by using machine learning. The machine learning learns the patterns in the data and predict future events or predict[28]. The LIBS data contain many features; in other words, it is also called as dimensions. Not all the dimensions are useful; some of the dimensions contain important information while other dimensions

contain less critical information. The principal component analysis (PCA) has been used for dimensionality reduction purpose. There are other techniques as well which could have been used for dimensionality reduction purposes such as wavelet transform (WT) and independent component analysis (ICA). However, PCA is a linear dimensionality reduction technique, and the data will be rearranged in decreasing order of variance. The dimensionality can be reduced manually, by choosing an important dimension which represents an important chemical element, and it includes human interference for pre-processing data. In order to keep it completely autonomous, the principal component analysis (PCA) algorithm has been used.

Machine learning algorithms are used for decision making purpose[28]. The machine learning algorithms became popular because most of the classification challenges contain high dimensionality data, a human can perceive only three dimensions, so the machine learning algorithms can build models using high dimensional data, these models can be integrated into working software support the kinds of product. Here, the LIBS data obtained from any material such as liquid, gas or solid contains patterns, these patterns reveal the information about the material, these patterns can be learnt by the machine learning algorithm, and it can predict the unknown inputs. In this work, three algorithms have been used, and those are Support Vector Machine (SVM), K-nearest neighbour search (KNN) and Artificial Neural Network (ANN) with backpropagation algorithm.

The rise of artificial intelligence has helped in deciding remote sensing. The NASA in its Mars rover mission has accomplished multielement chemical analysis on the mars planet using ChemCam device[29]. The ChemCam device is a LIBS device which is used

for multielement chemical analysis on the mars[30] [31]. Statistical and machine learning techniques achieve the atomisation in the multielement chemicals.

One of the challenges in the data obtained from the LIBS device is pre-processing of the data. The LIBS data is high dimensional; to reduce the dimensionality is a significant challenge. This can be achieved by PCA. It is a favourite feature extraction technique[32]. It projects highly dimensional data onto a lower dimension using a linear transformation [33], [34]. Typically, not all features are essential for creating a model, and there is a threshold number of dimension above which the performance of the model can degrade [35]. The wavelet transform is an alternative to PCA, and it is applied to the LIBS signals [36]. Wavelet transforms and PCA functions in a different way, wavelet transforms use wavelets derived from the data as a basis, and PCA uses an eigenfunction derived from the data. In [37], [38], the authors carried out the manual identification of features based on the peak of the critical spectra and the respective wavelength.

As NASA has developed ChemCam for Mars rover mission[39], on similar ground, The University of Tokyo has developed the ChemiCam device to probe the ocean[2]. Much work has been carried out by our team using ChemiCam. The ChemiCam equipped with ROV is used for deep-sea hydrothermal vent analysis and performed in-situ multielement chemical analysis[2]. Hydrothermal vents are essential since it contains essential industrial minerals and metals such as copper, lead zinc and iron[40]. The ChemiCam device has been inspected for generating spectra at high pressure up to 30MPa [41], [42]. This ChemiCam device is inspected for both liquids and immersed solids in in-depth ocean surveys [2]. A calibration-free technique was applied to measure the brass alloy standard

samples that were merged in seawater [42]. The LIBS spectra's nonlinear temperature is quantified using the principal component regression (PCR) and partial least squares (PLS) algorithms[43].

After developing the chemicam by the University of Tokyo and tested under various conditions, there is still one thing that was remained which can help in deciding remote sensing. The use of machine learning algorithms to create a model, the model which is trained from enough data, if it identifies the unknown data, then this will save much time. By using such a model, rock identification can be made in real time in the ocean. This was the research gap which was necessary to fill. This thesis is a successful attempt to fill this gap.

1.9 Motivation and research gap

In section 1.2 and 1.3, it has already been explained about the need for minerals, metals, gas and oil to the human society. The demand has been being rapidly increasing, and the available resources are not enough to meet the demand, in such condition, it better to search for other viable options[5]. Until now, we have done onshore mining, and it is estimated that the resources obtained from onshore mining may not meet future demand. It is known that earth contains only around 30% of land and 70% water. So it is evident that most of the energy resources are hidden in the ocean. Ocean mining also knew offshore mining. Other than onshore and offshore mining, there is one more option that is, outer planet mining, but in the situation, it seems to be very difficult for at least few more decades and also it will be much expensive than offshore mining.

This thesis has explored the use of machine learning algorithms with help laser-induced breakdown spectroscopy (LIBS) ChemiCam device to aid in the hydrothermal rock classification not only concerning its label but also concerning its group. The group of the rocks have been formed based on the copper, zinc and lead elemental content in the rocks. The outcome machine learning model of this thesis will be helpful for in-situ multi-element chemical analysis in real time as shown in Fig. 1.8

1.10 Thesis Overview

This thesis focus on classification and statistical analysis of rocks obtained from Okinawa and measured using laser-induced breakdown spectroscopy. In the Chapter 1 introduction, the need of the research, literature survey, motivation and research gap have been explained along with that, the significant keywords have been explained such as marine minerals, in-situ chemicals, LIBS, ChemiCam etc. Chapter 2 is focused on ChemiCam device, data collection during sea-trials and experimental setup during in the lab. Chapter 3 is focused on methodology, machine learning and statistical analysis methods have been explained. Chapter 4 focused on, results and analysis, the results obtained for label classification of 10 rocks and 30 rocks, and group classification of 30 rocks have been explained. The last chapter 5 is focused on the conclusion of the whole thesis.

1.11 Figures

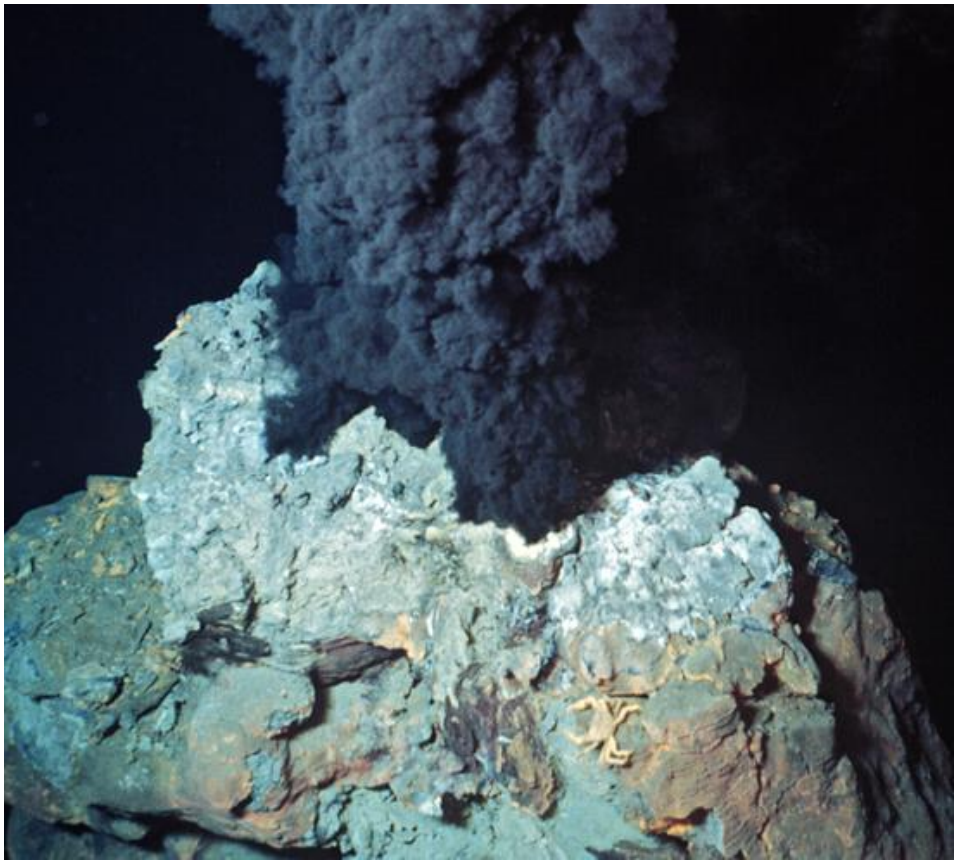


Fig. 1. 1 Active Hydrothermal vent. Image from NDSF, ROV Jason, © Woods Hole Oceanographic Institution

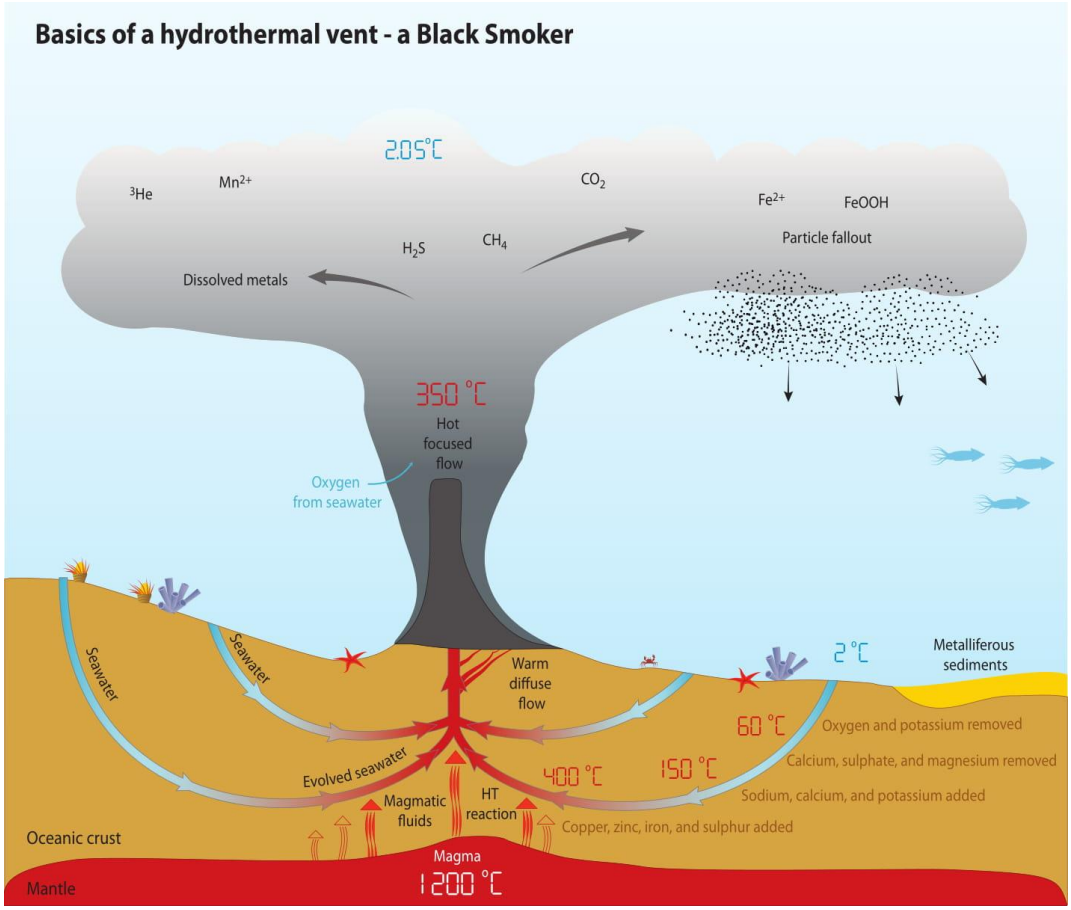


Fig. 1. 2 Basics of a Hydrothermal vent

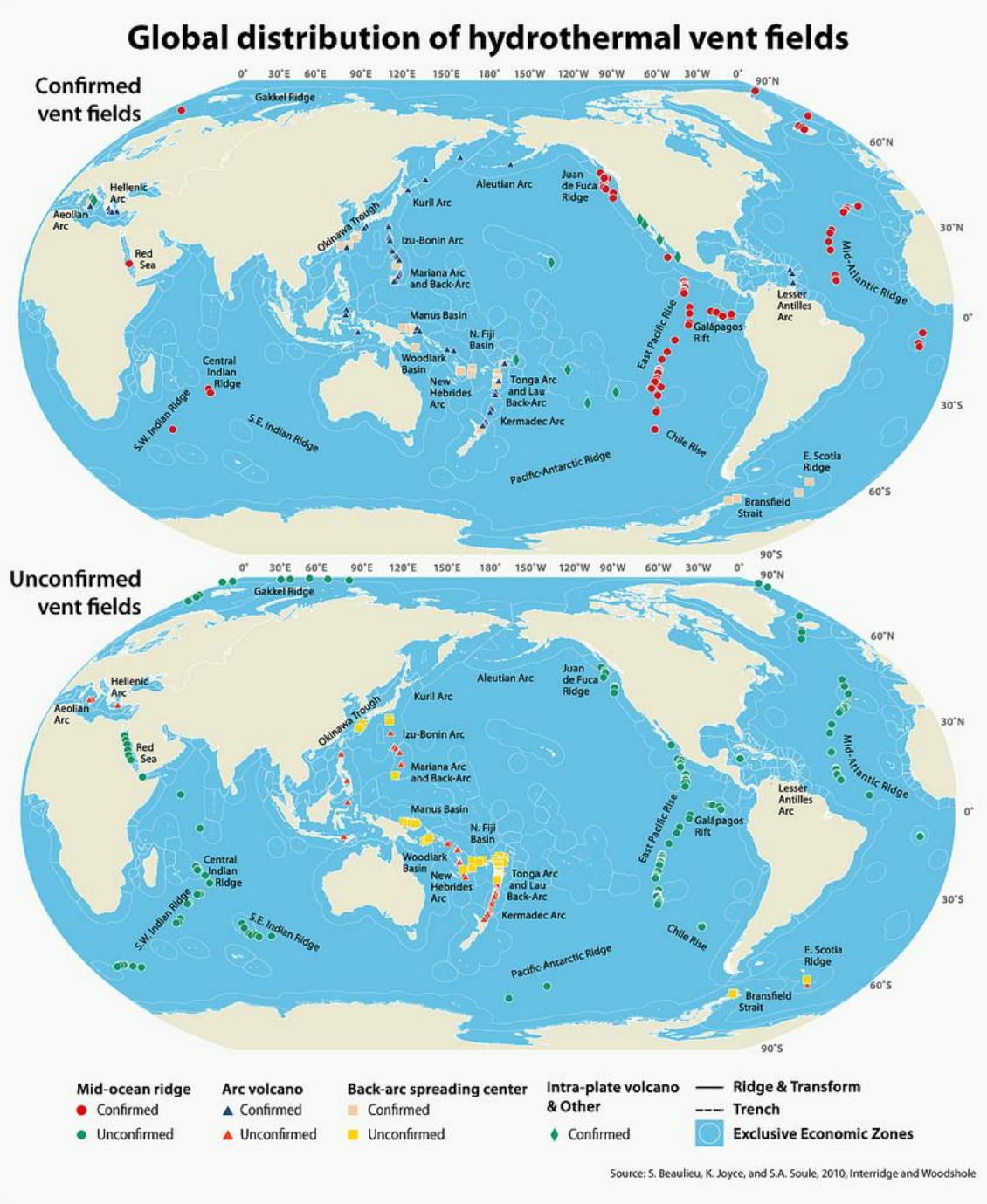


Fig. 1. 3 Distribution of Hydrothermal vents

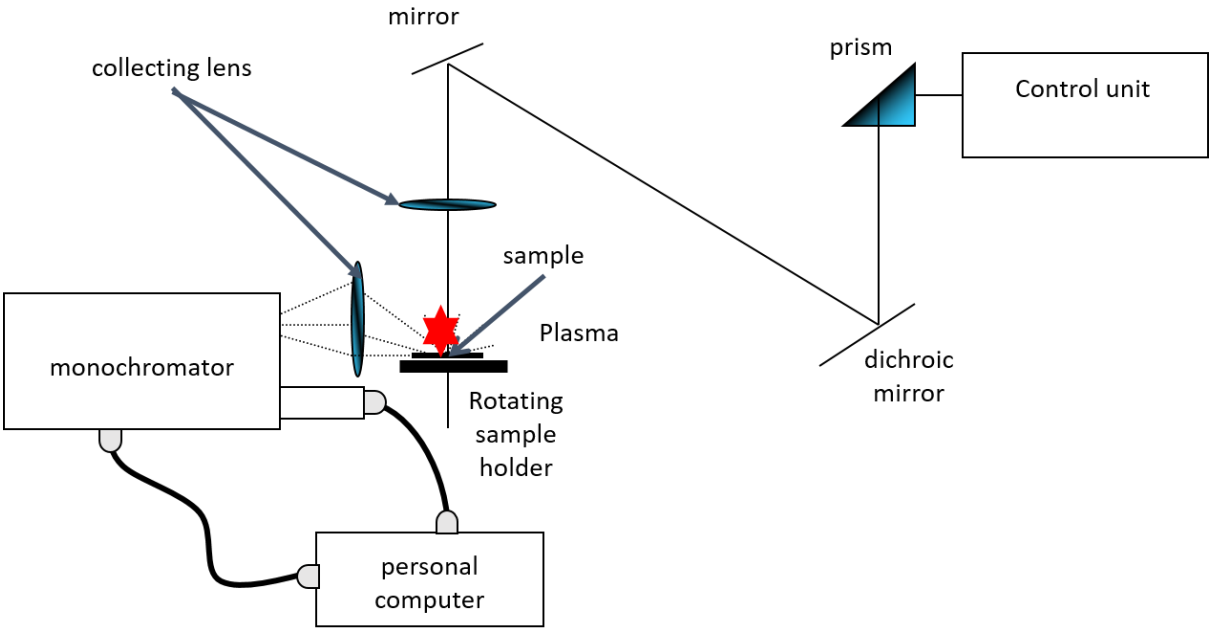


Fig. 1. 4 Working principle of LIBS

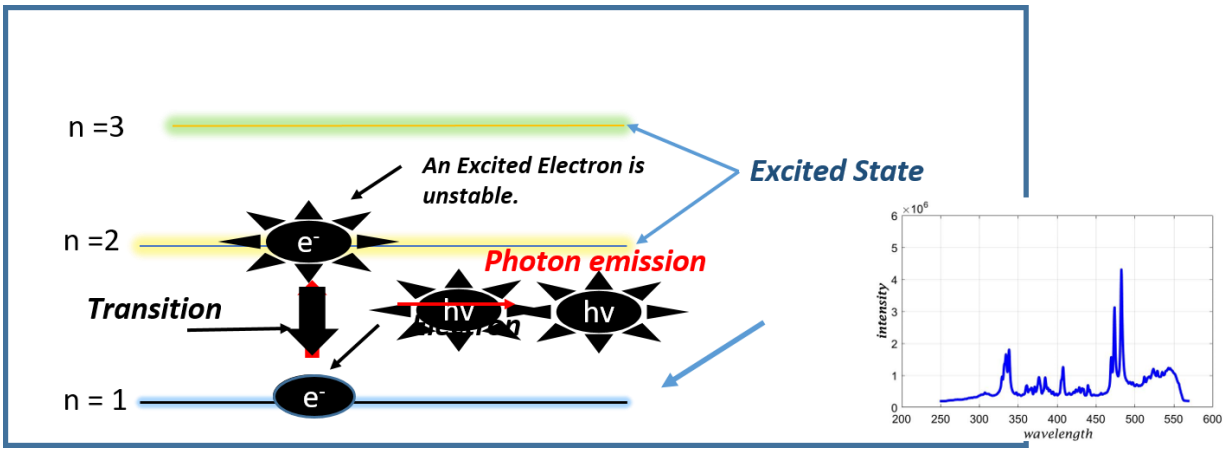


Fig. 1. 5 Plasma emission principle

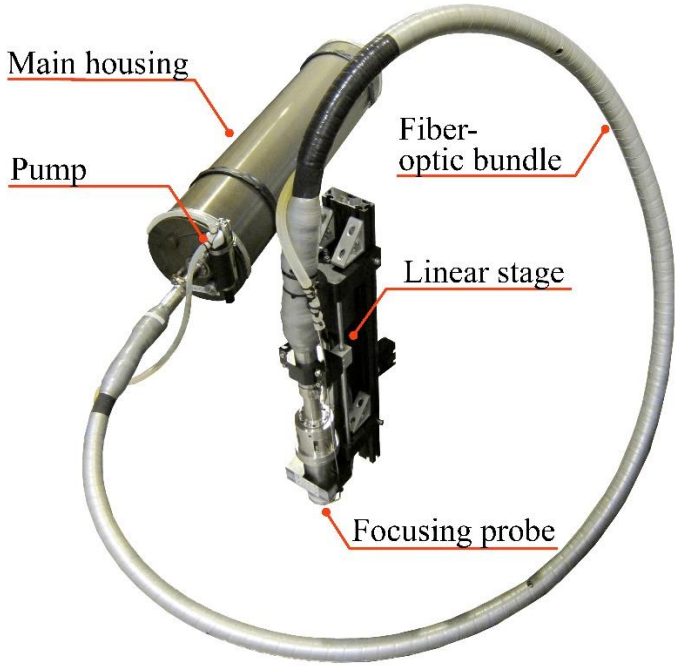


Fig. 1. 6 The 3000 m depth rated LIBS device ChemiCam

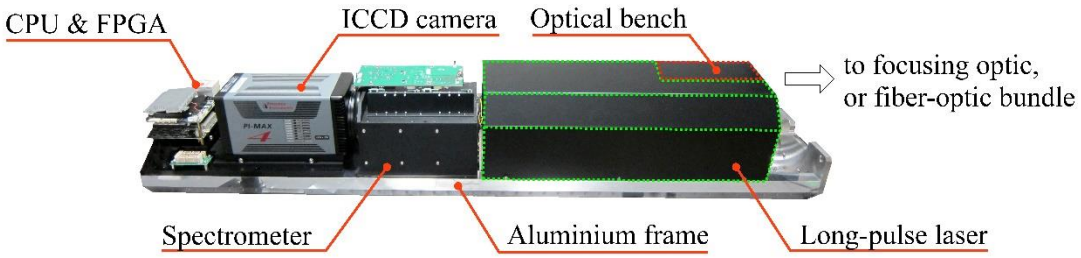


Fig. 1. 7 A long pulse laser beam

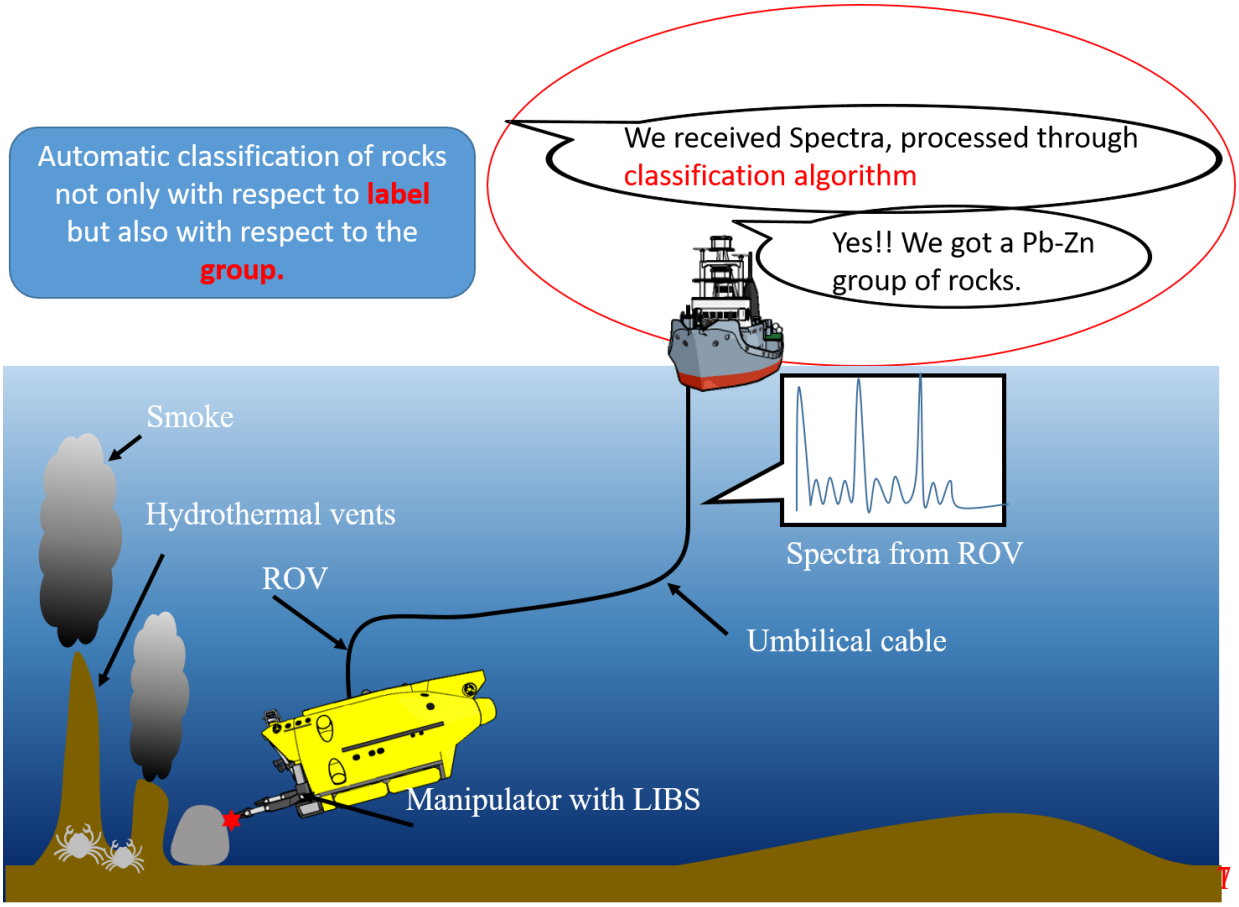


Fig. 1. 8 Research gap and motivation

1.12 Tables

Table 1. 1 Deep-sea Hydrothermal Vents

	Hydrothermal Vent
Depth [m]	700 ~ 2000
Mineral from	Polymetallic sulphides
Element	Cu, Pb,Zn,Au, Ag

Characteristics	Hot vent field with a maximum fluid temperature of 270 to 380 degree Celsius
Sources	Ocean Basins
Location	Back-arc basin, mid-ocean ridge, Volcanic Arc

Table 1. 2 Specifications of ChemiCam device

Specification	Quantity
Length (m)	1300
Diameter (mm)	300
Maximum depth (m)	3000
Weight in the air (kg)	160
Weight in water (kg)	40
Power consumption (W)	140
Power Supply (VAC)	100
Communication	RS 232 or Ethernet
Laser type	Q-switched DPSSL Nd: YAG
Laser pulse type	Long pulse
Maximum pulse energy (mJ)	20
Pulse duration (ns)	150
Laser wavelength (nm)	1064
Frequency (Hz)	2
Spectrometer type	Czerny-Tuner
Spectral range (nm)	320-550
Spectral resolution (nm)	0.25
Detector Type	ICCD
Number of pixels	1024 x 256

Fibre length (m)	5
Fibre type	1 600-µm-core and 45 100 µm-core bundle
Focussing lens	5 x Objective lens
Functions	Linear stage pump for removing debris
Autofocus	Yes

2 Machine Learning Models for Classification

2.1 Introduction

The search of ocean mining to meet the hunger of energy is on demand, the use of artificial intelligence is the smart way meet the hunger of energy. The hydrothermal deposits carry importance since it contains essential metals and minerals. The second chapter is focused on data preparation. Now in this chapter, how this data has been used to create the learning model which identifies the certain patterns from the data and predicts the unknown data, the methodology has been explained, the algorithms used for pre-processing, classification and statistical analysis purpose. The data was normalised before it pre-processed. The new technique has been investigated that is "linear detrend" on the spectroscopic data to see the effect of it on the classification results. The principal component analysis has been used as a dimensionality reduction technique.

The overview of the methodology is shown in Fig. 2.1. Total of 300 LIBS shots was recorded on each rock. This dataset is divided into two parts, in the first part, only 100 shots of each rock were considered, and in the second part, all data that is 300 shots of each rock were considered as shown in Fig. 2.1. The dataset has been divided into two parts, to investigate the effect of increasing the dataset on the learning models.

Further, the effect of linear detrend has been investigated. To investigate the effect of linear detrend, each part of the dataset is verified by applying linear detrend operation and without

linear detrend operation. So the total four learning models have been verified as shown in Fig.2.1

2.2 Pre-processing of data

The Fig.2.2 shows the sequence of methods followed to create the model. The final block shown in Fig 2.2 shows the statistical analysis using the Wilcoxon signed rank test. The normalisation process brings the feature values between -1 to 1. This is achieved by subtracting the mean value and divided it by standard deviation from each feature. The linear detrend is removes the regular shift occurred due to electronic noise. PCA is a feature reduction technique since all feature does not carry important function. Further various classification algorithms applied such as SVM, KNN and ANN to classify each spectrum. The details of each block are given further explanation.

2.2.1 Normalisation

The normalisation process brings all features on the same scale with a mean of -1 and a variance of 1 [44]. The normalisation process is achieved by subtracting the mean of the feature and divided it by the standard deviation of the feature.

$$X' = \frac{(X - \bar{X})}{\sigma} \tag{2.1}$$

Where X is the original feature, $\bar{X}$ is a mean value, σ is the tandard deviation, X' is a normalized value

2.2.2 Principal component analysis

The Principal component analysis is a dimensionality reduction technique. The data produced from the laser-induced breakdown spectroscopy, i.e. ChemiCam device is 1024 dimension. As the attributes of the data increases, the complexity of the data also increases, most of the time, all dimensions do not carry relevant information, in such cases, the dimensionality reduction technique plays a vital role to discard the redundant information from the data[45]. The principal component analysis (PCA) is one of such techniques.

- **Definition of Principal Components**[46]:

Suppose x is a vector of r random variables and x^T denotes the transpose of x. So,

$$x = [x_1, x_2, x_3, x_4, \dots., x_r]^T \tag{2.2}$$

Let us look at the linear function $\propto_1^T x$ of the elements of x which has maximum variance, where, $\propto_1$ is a vector of r constants, $\propto_{11}, \propto_{12}, \dots, \propto_{1r}$, so that

$$\propto_1^T x = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1r}x_r = \sum_{j=1}^{r} \propto_{1j} x_j \tag{2.3}$$

So the aim is to find out the linear function that transforms random variables into a new random variable so that new variable $\propto_1^T x$ has maximum variation.

Next step is to find out a new set of random variables using linear functions based decreasing order maximum variance, these new random variables are called as principal components.

2.3 Machine learning

2.3.1 What is Machine Learning

Today is a big data era, there are about 1 trillion webpages and almost one hour of video is being uploaded for every second[47]. Similarly, data is being been generated everywhere and every point for example in medical labs, in the ocean from Hubble telescope and the images sent by Mars rover etc., the meaningful information can be obtained from this big data. Machine learning is a set of methods which uncover meaningful information from the big data. Machine learning is a subtopic of artificial intelligence (AI). It provides the ability to learn and improve the experience automatically without being programmed. Its main aim is to allow a computer to learn automatically without human assistance. In machine learning, the learning occurs by extracting useful information from the data through algorithms that distinguish between signal and noise, once it finds the usage pattern, it lefts the everything else as noise, for this reason, the machine learning algorithms also known as pattern recognition[48]. Learning requires intelligence, and it covers a wide range of processes, so it is difficult to define precisely[47]. Zoologist and psychologists study learning in humans and animals, and with the help of this, we can design a machine which learns similarly. One of such learning is artificial neural networks, similar to biological learning.

2.3.2 Types of Machine Learning Methods

Machine learning algorithms are roughly categorised into two main types. Those are supervised learning and unsupervised learning.

2.3.2.1 Supervised Machine Learning

In this type of algorithms, labelled data is used to predict future events. It maps the input variable and the output variable. Supervised learning algorithms are used when we know the exact label of output data. The objective is to learn a mapping from inputs x to outputs y, given a labelled set of input-output pairs $M = \{(x_i, y_i)\}_{i=1}^{N}$ here M is called the training set and N is the number of training examples. In the same way, the response variable y_i can be anything, in general $y_i \in \{1,2, \dots, C\}$. The response variable can be categorical or nominal variable from some finite set. If C=2 then this is called binary classification if C>2 then this is called as Multiclass classification.

Some of the examples of supervised learning are as follows,

- Predicting the stock market
- Predicting age viewer of youtube
- Predicting the location of 3d space of robot arm end effector.
- Predicting temperature at any location

2.3.2.2 Unsupervised Machine learning

In this case, only output data will be given without any inputs; the goal is to discover the "interesting patterns" in the data; this is also known as "knowledge discovery". In

supervised learning, the desired output is known, but in this case, it is unknown. This kind of algorithms is used when one does not know corresponding output variables concerning input data. The goal of this kind of algorithms is to find out the clustering or structure or association among the data. There are types of unsupervised learning such as clustering and data compression. Clustering algorithms run through data and find the natural clusters if they exist. There are various types of clustering algorithms such as K-Means clustering, Hierarchical clustering and Probabilistic clustering. There are some challenges in unsupervised learning algorithms for example in supervised learning algorithms, the label of the data is given so we can measure how accurately the algorithm is working, but in the case of the unsupervised learning algorithm, it is difficult to know how better the algorithm is working.

2.3.2.3 Semi-supervised Machine learning algorithms

The difference between supervised and unsupervised machine learning algorithms is that supervised learning dataset has labels whereas unsupervised learning algorithm dataset has no labels. The semi-supervised learning algorithms are a combination of labelled and unlabeled data. Labelling a massive amount of data is time-consuming and expensive[49]. This means a lot of unsupervised data combined with supervised data, tend to reduce the cost spent on building the model. The applications of semi-supervised learning are like webpage classification, speech recognition and genetic sequence.

2.3.3 Support Vector Machine (SVM)

The Support vector machine learning algorithm is a supervised learning algorithm; it defines a hyperplane that divides the data in descriptive space forming a group of data[50]. SVM is a kernel-based learning algorithm, where the kernel is a mapping function. SVM transforms given dimensional space into higher dimensional space. It transforms into higher dimensional space using the kernel trick described below. The SVM method introduces the trade-off parameter referred to as 'C' that penalises the data points that cannot be separated. The user can choose the kernel 'K'. If there are two points M1 and M2 in space, then the chosen 'K' kernel maps these points in the transformed space, as shown in Fig. 2.3

$$S(M_1, M_2) = \langle K(M_1), K(M_2) \rangle \tag{2.4}$$

The kernel trick defines the similarity function 'S'. There are two types of kernels used, namely, the SVM linear kernel (SVM Linear) eq. (4) moreover, the SVM radial basis function (SVM-RBF) eq. (5).

$$S(M_1, M_2) = M_1^T . M_2 \tag{2.5}$$

$$S(M_1, M_2) = e^{(-\gamma \| M_1 - M_2 \|^2)} \tag{2.6}$$

Where γ defines how far the single training sample affects the performance. This parameter is inversely proportional to the influence of samples selected by the model as support vectors.

2.3.4 K-Nearest Neighbour Search (KNN)

The KNN algorithm is widely used for supervised classification problems where the label of the data is known [51]. In the KNN algorithms, the data points are spread in the metric space. The Euclidian distance is used to determine the class of the test data point based on the most significant number of k-closest training data points [52].

It falls in the supervised learning family. Suppose training observations represented as (x, y). To find out the relationship between x and y, the goal is to learn a function $h: X \rightarrow Y$, so that test observations x, h(x) can correctly predict the corresponding output y.

KNN is a non-parametric and instance-based learning algorithm. Non-parametric methods that do not make explicit assumptions about the data, avoiding the probability of misleading underlying distribution of the data.

The KNN works by forming a majority vote between the K most similar instances to a given "unseen" observation. The distance metric between two data points is given as,

$$d(x, x') = \sqrt{(x_1 - x_1')^2 + (x_2 - x_2')^2 + \cdots + (x_n - x_n')^2} \qquad (2.7)$$

2.3.5 Artificial Neural Network (ANN)

Artificial neural networks are one of the vital technique used in machine learning. It is a brain-inspired system which replicates the way human learns. The neural network consists of the input layer, an output layer and hidden layers as shown in Fig. 2.5. If the numbers of hidden layer are one, then it is called a shallow neural network, if the number of hidden layers is more than one, then it is called a deep neural network. It is an excellent tool to find out patterns which are far too complicated. Neural networks, also known as perceptron concepts exists since the 1940s, but since the last several decades, they have become a significant part of artificial intelligence. In the ANN algorithm, a multilayer perceptron with a tangent hyperbolic sigmoid transfer function is used. The network has one hidden layer with 'h' hidden neurons, and the final layer is the output layer[53]. The neural network is trained using a feedforward algorithm [54].

2.3.6 Cross-validation technique

The cross-validation (CV) technique is used to find out the best-optimised parameters. It finds out the best parameters by generalising across all the samples. This technique assists in selecting the best model, that does not overfit. The dataset is divided into training and test datasets with a ratio of 0.8 and 0.2. Further, to find the best parameter of the algorithms, the training dataset is divided into n folds. The $(n-1)$ fold datasets are used to train the model and are evaluated on the remaining fold of the data; this process is repeated n times, and the average CV hit ratio is calculated. The parameters that provide the highest CV hit

ratio are used to create the training model and are evaluated on the test dataset. The flow chart is shown in Fig. 2.6 [55] As ma easure of the performance, the hit ratio of the classifier on the test dataset is defined as,

$$hr = \frac{T_{hit}}{T_n} \tag{2.8}$$

Where, hr hit ratio, T_{hit} total number of correctly classified samples, T_n total number of samples

2.4 Statistical Analysis

2.4.1 Wilcoxon Signed Rank Test

The test is named for Frank Wilcoxen who proposed this technique[56]. Sidney Siegel popularised this test in his book on nonparametric statistics[57]. It is a non-parametric statistical test used to compare two related samples. So the statistical significance is evaluated for the algorithms by using a one-sided Wilcoxon signed rank sum test [58]. This test is a nonparametric method. It does not assume that the population has any particular form, unlike parametric tests such as the t-test or the analysis of variance (ANOVA). One of the principal reasons to choose a nonparametric test such as the Wilcoxen signed rank test is that it works well with small sample sizes and has few assumptions about the data. Some of the assumptions such as data are paired and come from the same population. Each pair is chosen randomly and independently.

2.4.1.1 The test procedure of Wilcoxon signed rank test[59]

Let "N" be the total number of pairs. Thus there are total 2N data points. For pairs, $i = 1,2,\dots,N$, let $x_{1,i}$ and $x_{2,i}$ denote the measurements.

Let us make two hypothesis

H1- the difference between the pairs follows a symmetric distribution around zero.

H2- the difference between the pairs does not follow a symmetric distribution around zero.

1. For $i = 1,2,\dots,N$, calculate $\left|x_{2,i} - x_{1,i}\right|$ and $sgn\left(x_{2,i} - x_{1,i}\right)$, where sgn is the sign function.
2. Exclude pairs with $\left|x_{2,i} - x_{1,i}\right|$=0. Let N_r be the reduced sample size.
3. Order the remaining N_r pairs from smallest absolute difference to lathe rgest absolute difference, $\left|x_{2,i} - x_{1,i}\right|$.
4. Starting from 1, Rank the pairs. Ties receive a rank equal to the average of the ranks they span. Let R_i denote the rank.
5. Calculate the test statistic W

$$W = \sum_{i=1}^{Nr}\left[sgn\left(x_{2,i} - x_{1,i}\right).R_i\right] \tag{2.9}$$

6. Under the null hypothesis, "W" follows a specific distribution with no simple expression. This distribution has an expected value of 0 and variance of

$$\frac{N_r(N_r+1)(2N_r+1)}{6} \tag{2.10}$$

"W" is compared to a critical value from a reference value from a reference table. The two-sided test consists in rejecting H_0 if $|W| > W_{criticle,Nr}$.

7. As N_r increases, the sampling distribution of W converges to a normal distribution. Thus for N_r>20, a z-score can be calculated as $z = \frac{W}{\sigma w}$, where,

$$\sigma w = \sqrt{\frac{N_r(N_r+1)(2N_r+1)}{6}} \tag{2.11}$$

8. For N_r<20 the original test using the T statistics is applied.

So, this is the test procedure applied to the calculating significance of Wilcoxon signed rank test.

2.5 Summary

This chapter was focused on methodology. The explanation starts with why we need data processing and the necessary reasons for pr-processing. Further, the normalisation process has been explained. Linear detrend is a new technique for pre-processing; the necessary reasons to use this technique has been explained. The machine learning algorithms such as Support Vector Machine with two kernels those are a linear kernel, and radial basis function kernel has been explained. The other machine learning algorithms such as

Artificial neural network and K-nearest neighbour search has been explained. The statistical tool such as Wilcoxen signed rank test has explained which is used for comparing the performance of classification algorithms.

In the next chapter, I have explained how these algorithms are used on spectroscopy data to classify the spectra of each rock. Not only concerning labels but also concerning its geological group. The detailed analysis of results is given.

2.6 Figures

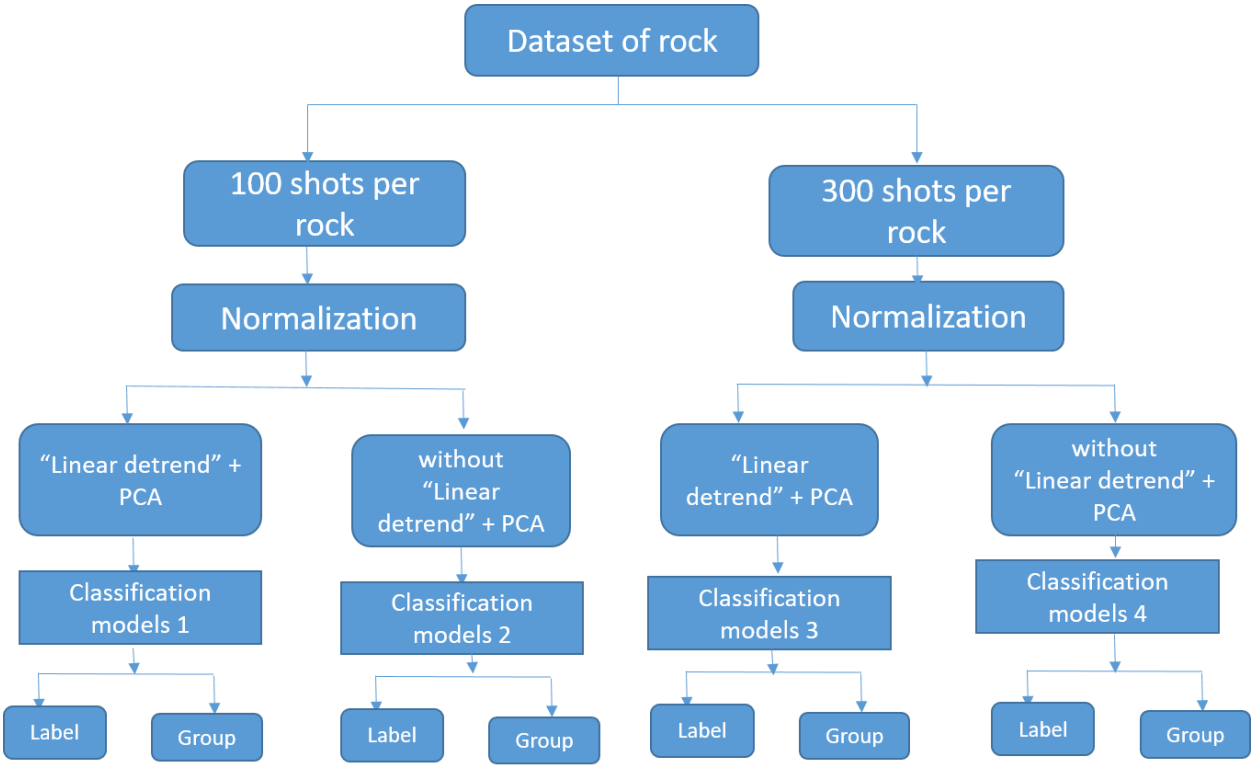


Fig. 2. 1 Methodology Overview

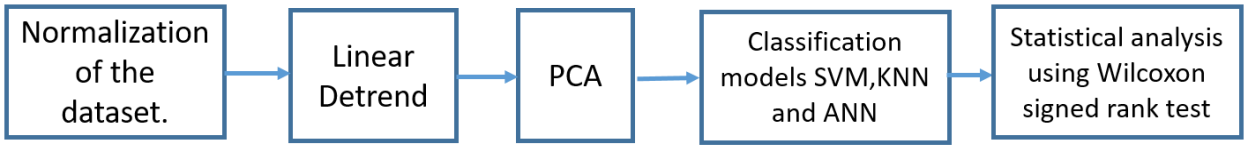


Fig. 2. 2 Methodology sequence

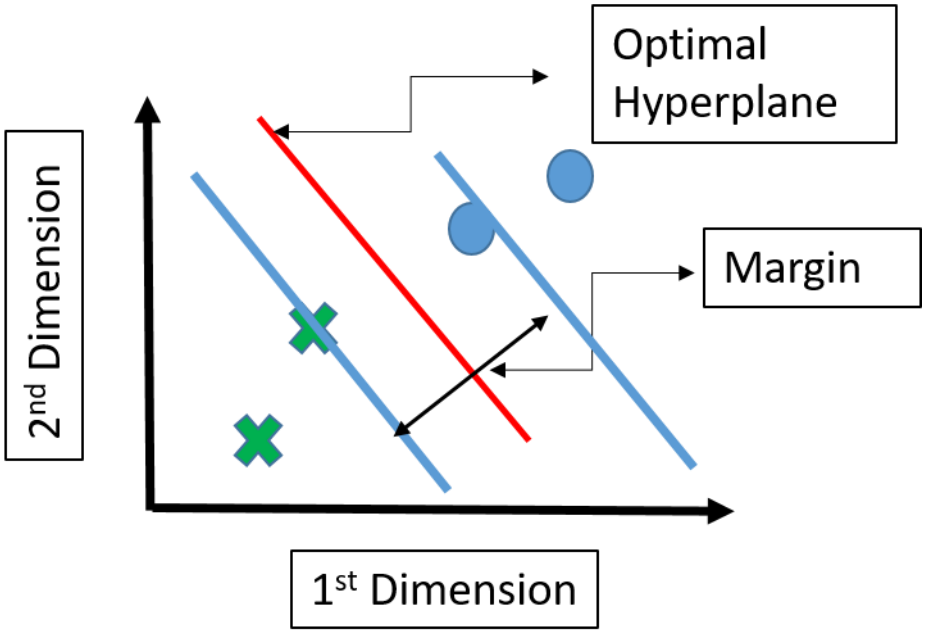


Fig. 2. 3 2D-view of SVM classification

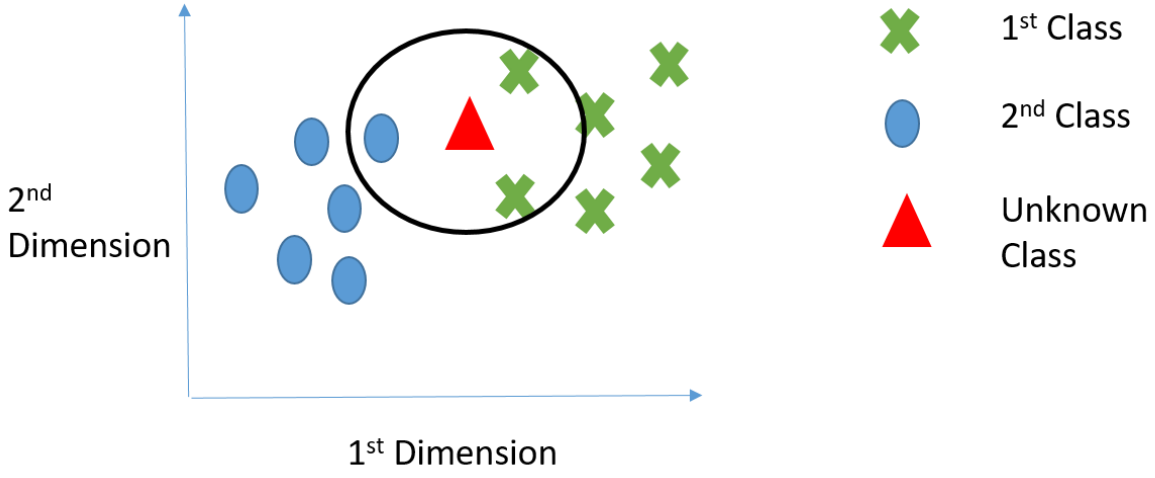


Fig. 2. 4 KNN classification in 2-Dimension

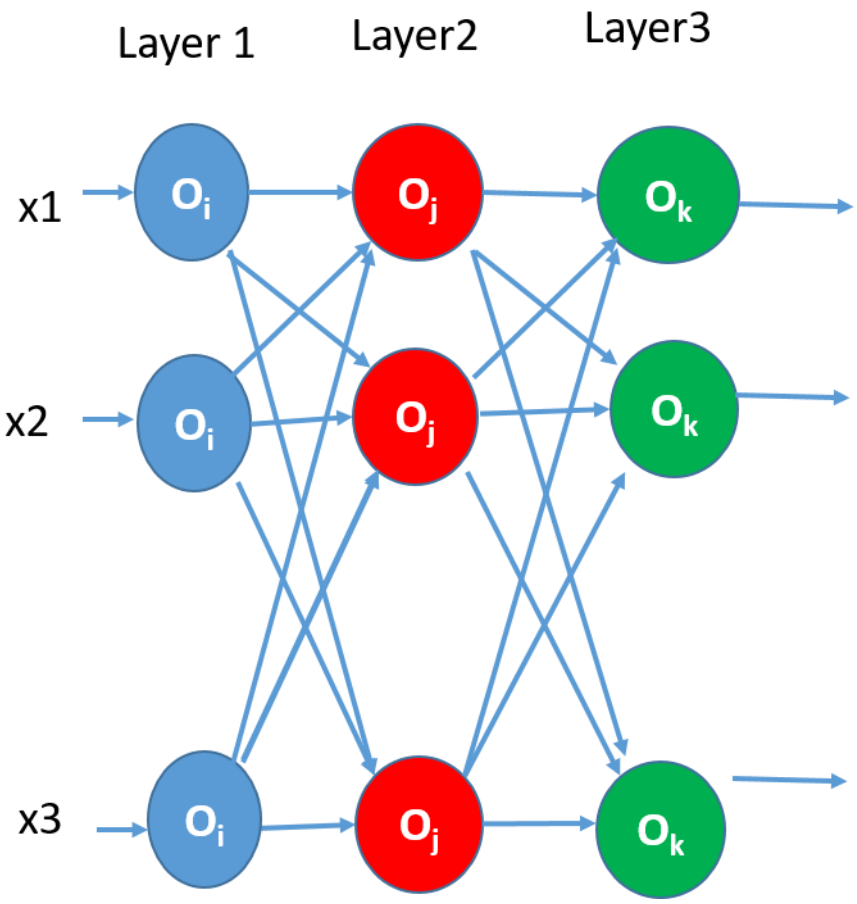


Fig. 2. 5 A backpropagation neural network

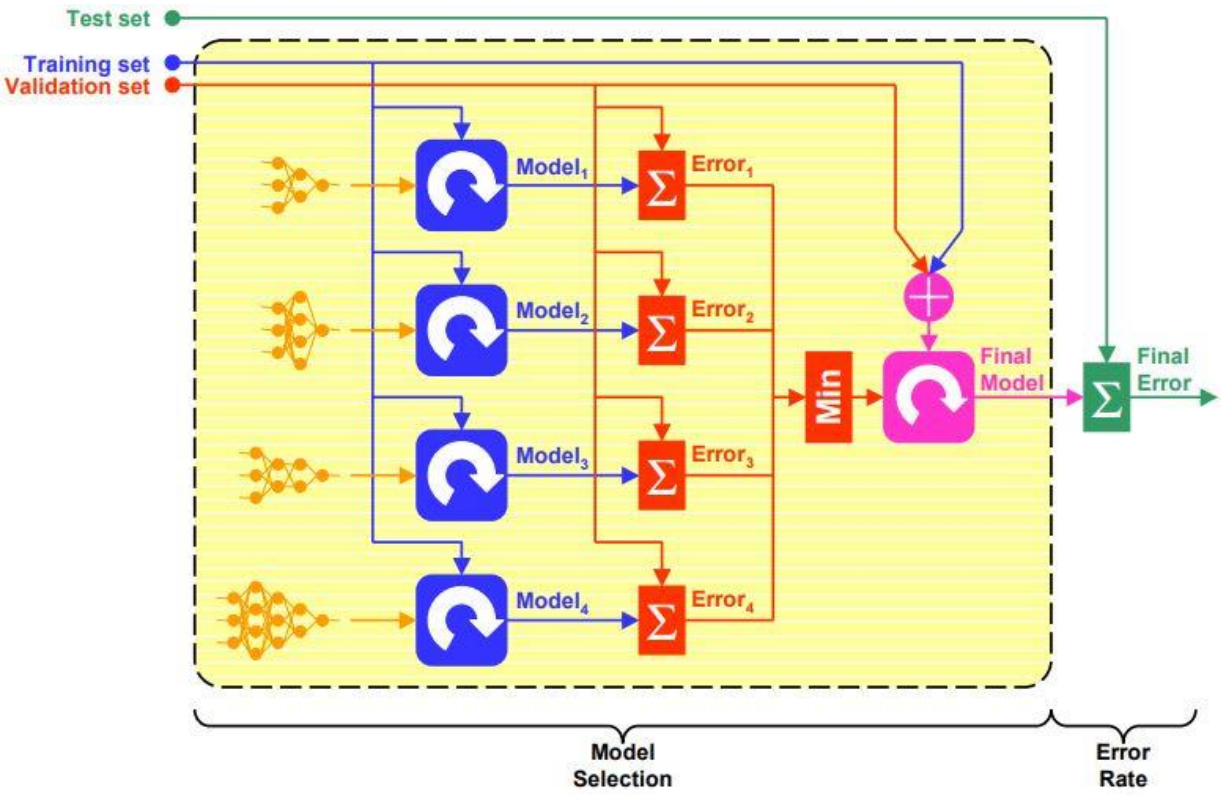


Fig. 2. 6 Flow chart of cross-validation algorithm

3 Result and Discussion

The first chapter was focussed on introduction, where the background of research, literature survey, research gap and motivations have been explained. The second chapter was focussed on the machine learning algorithms used in this thesis. In this chapter, we focused on LIBS experimental setup to get the rock dataset, the geological characteristics of the data, pre-processing of the data, training a model using machine learning algorithms and testing the performance of the trained algorithm using test data, the statistical analysis of the performance of the classifiers.

3.1 Experimental Setup

The experimental setup is shown in Fig. 3.1[53]. A pulse is generated with a 1064 nm Nd: YAG Q-switched laser having the energy of the laser pulse of 5 mJ. The pulse duration is 250 ns. The frequency is 2 Hz. The laser pulses are struck onto the rock pellets using 600 μm fused silica fibre. The high-power laser pulse produces a plasma on the rock pellet. The plasma is the fourth state of matter; other states are solids, liquids and gases. It consists of ions, atoms and charged particles. A plasma produces a variety of the wavelength of light that is collected through a spectrograph using an intensified charge coupled device (ICCD) records all the wavelengths simultaneously after the diffraction grating has dispersed the light. The quantitative analysis of plasma is made possible because the wavelength emitted from rock pellets unique to each species. The maximum signal-to-noise ratio was achieved with a gate delay and a width of 400 ns and 500 ns, respectively. A Czerny-Turner type

spectrometer is used with an input slit measuring 20 × 800 μm, the spectral range is from 320nm to 550nm, and a spectral resolution varies from 0.22nm to 0.24nm across the range.[60].

Each laser shot creates one vector of 1024 dimension. Each dimension corresponds to a specific wavelength of light. Total of 30 rocks have been used, and 300 laser shots have been recorded. Fig. 3.2 shows the matrix of LIBS data.

3.2 Classification model

Three types of classification models have been created those are support vector machine (SVM), k-nearest neighbour search (KNN) and artificial neural network (ANN). These are the supervised classification algorithms because the label assigned to each data is known. A label classifies the signal to each rock in the dataset [61]. The data is described as (x_i, y_i), where x_i is a k-dimensinal input vector containing k values (i.e., attributes or features), $i = 1, \ldots, N$ represent rocks and $y_i \in M = \{1,2, \ldots, m\}$ is a class label from the finite set M containing m classes[62]. The goal of classification is to learn a function f that assigns a class label from the finite set. The Fig. 3.3 shows the model training using machine learning algorithms.

3.3 Rock label classification

To verify the effect of the "detrend" operation on the data and the effect of varying dataset size on the classification models, the dataset is divided into 4 cases as explained in section 2.1. For the first case, the 100 shots of each rock were considered without the detrend operation, and in the second case, 100 shots of each rock were considered with the detrend

operation. In the third and fourth cases, 300 shots were considered without detrending and with detrend, respectively.

Fig. 3.4[53] shows the difference between the LIBS spectra before and after the removal of the linear trend (data cleaning). The PCA algorithm is applied to both cases, to investigate the effect of dimensionality reduction. The principal components are extracted after the normalisation operation, and these principal components are the eigenvalues of the attributes correlation matrix. These are arranged in descending order. Fig. 3.5[53] shows the scatter plot of the first three principal components of the data. Table 3.5 shows the percentage of variance of the first three principal components with and without detrend. It shows that without detrend operation, the percentage of variance of the first principal component is 70% whereas without detrend operation is 31%. It implies that PCA with detrending increases the variability among the rocks. Therefore it increases the classification sensitivity as shown in Fig. 3.6[53] and Table 3.3 [53]. Table 3.1 [53] shows the various tuning parameters. These parameters were varied over a broad range, to determine the values that are given in an optimised training model. The grid search technique is used to find the optimised parameters [63]. From Table 3.2 [53], the number of principal components (N_{PC}) is different for all of the algorithms. The other parameters were such as the regularisation parameter 'C' in case of SVM linear and 'C' & 'γ' in the case of SVM-RBF. The number of nearest neighbors in the case of the KNN algorithm and number of hidden layers in case of the ANN algorithm in the CV loop The SVM linear, SVM-RBF, KNN and ANN classification techniques are explained in the methodology section.

Table 3.3 [53] shows the average sensitivity, specificity and accuracy of each algorithm for every set of data. Usually, the sensitivity and specificity are calculated for binary classes, but in our case, the data are multiclass, so the one vs all approach is used to calculate the sensitivity and specificity[53]. The sensitivity, specificity and accuracy of the classifiers are calculated using the formula given in equation 3.1,3.2 and 3.3 respectively.

$$specificity = \frac{TN}{TN + FP} \tag{3.1}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{3.2}$$

$$sensitivity = \frac{TP}{TP + FN} \tag{3.3}$$

Where, TP: True Positives, TN: True Negatives, FN: False Negatives, FP: False Positives

The sensitivity measures the percentage of accurately identified spectra; for example, the percentage in the rock two spectra is correctly identified as rock 2. Fig. 3.6[53] shows a graph of all the combined dataset of all used algorithm vs average sensitivity. It can be viewed that there is an increase in sensitivity when a detrend is performed. There is an increase in classification sensitivity when the 100 shots per rock dataset size are used with the detrend operation compared with the without detrend operation. For example, the ANN sensitivity increased from 68% to 74.6% when the detrend operation was operated on the dataset of 100 shots. In the case of the 300 shots per rock dataset size, there is a small increase in sensitivity when the detrend operation is performed. For example, the

sensitivity of ANN is increased from 87.5% to 89.5%. It can be concluded from Fig. 3.6[53] that the detrend operation is a vital operation for increasing the true positive rate. The classification models can efficiently learn the patterns in the data when the 300 shots per rock dataset are used. Therefore there is no much effect of linear detrend on a relatively larger dataset.

The specificity is a measure of the true negative rate; for example, the percentages of not rock one spectra are correctly identified[53]. The specificity of every the algorithm is approximately 99%. It shows that all the algorithms have a better ability to identify true negative results. The accuracy of the classifier is the ratio of the combination of true positive rate and true negative rate to the total population. There is a slight improvement in accuracy as the effect of detrending and dataset size. For example, it can be seen from Table 3.3 [53] that ANN accuracy increases from 97.8% to 98.3% when detrend operation is performed on the 100 shots dataset. In case 300 shots dataset the accuracy is increased from 99.3 to 99.4 in case of ANN. From Table 3.3 [53] it can be seen that SVM linear algorithm has performed best among all algorithm, it shows that, data are linearly separable at higher dimensions. The ANN algorithm has performed better than KNN as shown in Table 3.3 [53]

It can be inferred from the results that N_{PC} value differs from one algorithm to another and therefore the percentage of information from the original data also differs as shown Table 3.2[53]. Table 3.2[53] also shows that a number of required number principal components for ANN are much fewer than the SVM and hence the percentage of information from the original data is also less in the case of ANN compared to SVM. Hence it shows that the

ANN algorithm is computationally efficient in extracting useful information since it requires a minimum number of features to classify the rocks. The KNN is a nonparametric classification technique, so it demonstrated the lowest performance of the methods tested. The statistical significance is evaluated for the algorithms by using a one-sided Wilcoxon signed rank sum test [58]. This test is a nonparametric method. It does not assume that the population has any particular form whereas parametric tests such as the t-test or the analysis of variance (ANOVA). One of the principal reasons to adopt a nonparametric test such as the Wilcoxen signed rank test is that it works well with relatively smaller sample sizes and has few assumptions about the data. In the one-sided Wilcoxon signed rank test, the null hypothesis is that algorithm A and algorithm B have equal importance, and the hypothesis fails if the p-value is less than 0.05. Table 3.4[53] shows the Wilcoxon test applied to SVM-RBF, KNN and ANN against the SVM linear kernel since the SVM linear technique of 300 shots per rock with the detrend dataset has performed the best with the highest classification sensitivity. In the case of the RBF kernel, the test passes the null hypothesis since the p-value is higher than 0.05. In the case of the KNN and ANN algorithms, the p-value is less than 0.05, so the null hypothesis fails. From the above statistical analysis, it can be concluded that SVM linear and SVM-RBF perform equally well, whereas ANN and KNN perform poorly.

3.4 Rock Materials and geological classification

The composition of the 30 rocks calculated from ICP-MS data is shown in Table 3.6[53] The concentrations of Cu, Pb and Zn for each rock are plotted on a ternary diagram as

shown in fig.3.7[53]. The boundaries for rock classification were defined by the C rate (CR) and Z rate (ZR), as shown in Eq. (1) moreover, eq. (2) [3].

$$C = \frac{Cu}{Cu+Zn} \times 100 \quad (3.4)$$

$$Z = \frac{Zn}{Zn+Pb} \times 100 \quad (3.5)$$

The boundaries in the ternary diagram are defined as follows[3]:

1) Pb-Zn deposits (Z < 60);

2) Zn-Pb-Cu deposits (C < 60, 60 < Z < 90);

3) Zn-Cu deposits (C < 60, Z > 90); and

4) Cu deposits (C > 60, Z > 60).

None of the rocks was classified into the Cu group.

The spectral wavelength identification can be done using website atomic spectra database of NIST[64]. The Fig.3.8[53] shows a spectrum of each group. The crucial elements such as Cu, Pb and Zn wavelength lines were identified using the NIST website. The wavelength lines in nanometre are listed in Table 3.7[53]

For example, for Pb-Zn group's rock spectrum, the Cu wavelength line can be observed at 521.8nm and 515.3nm. Similarly, for Zn-Cu group's rock spectrum the Pb wavelength line can be observed at 404.9nm.

3.5 Rock Group Classification

The geological groups of the rocks are defined based on the ratio of Cu-Pb-Zn in each rock using the ternary diagram as shown in Fig. 3.7[53]. Here, the performance of the classifier is investigated while classifying the spectrum of each rock by considering its group. The

results of the rock group are derived from the rock label learning models results. So it is essential to understand how it is derived.

3.5.1 Rock group vs rock label

To understand the difference between rock label and rock group, it is important to go through the Fig. 3.9. This thesis aims to classify the rocks not only according to label but also concerning the group; all the rocks have been divided into three groups. Those are Pb-Zn, Zn-Cu and Pb-Zn-Cu groups. To understand the concept, consider the Fig 3.10[53] and take only two groups those are Pb-Zn and Zn-Cu. For example, rock 6 and rock 20 belong to Pb-Zn group, and rock 4 and rock 18 belong to the Zn-Cu group.

If the rock six is classified as rock six, then the rock group and rock label are the same. If rock six is classified as rock 20, then the rock label classification failed, but the rock group classification gets passed because rock 6 and rock 20 belong to the same group. Similarly, if rock six is classified as rock four, then rock group and rock label both get failed because rock 6 and rock 4 are of course two different label rocks but also belong to two different groups. In this way, the rock group and rock and label were classified.

Fig. 3.10 shows an effect of the increasing number of shots per rock on the accuracy of the rock group classification and detrend operation of each case. Fig. 3.10[53] and Table 3.8 [53] shows the 300 shots per rock dataset exhibit an increased accuracy compared to that of the 100 shots per rock dataset, and when the using the detrend operation with the 100 shots per rock dataset, a significant improvement in the rock group classification accuracy is demonstrated.

The one-sided Wilcoxon signed rank test technique is applied to search the statistical significance of the algorithms in the case of the 300 shots per rock with detrending dataset since it has performed well compared with the other methods [58]. The one-sided Wilcoxon signed rank test is applied concerning SVM linear. The result of the one-sided Wilcoxon signed rank test as shown in Table 3.9 [53] states that SVM linear and SVM-RBF methods have statistical significance over the ANN and KNN methods.

3.6 Summary

In this chapter, in the first part, under the title of rock label classification, the results of the SVM algorithm on ten rocks has been explained. The effect of the varying number of principal components on the accuracy of the SVM algorithms has been explained. Later, the number of rocks has been extended to 30 and along with SVM, the KNN and ANN algorithms have been used. Along with it, the effect of linear detrend operation on the classification sensitivity has been studied.

In the second part, the rock group classification has been explained. The analysis of each classification algorithm has been verified concerning its group. The performances of the algorithms concerning labels and groups have been shown graphically which makes it easier to conclude.

In the next chapter, the conclusion of the whole thesis has been drawn.

3.7 Figures

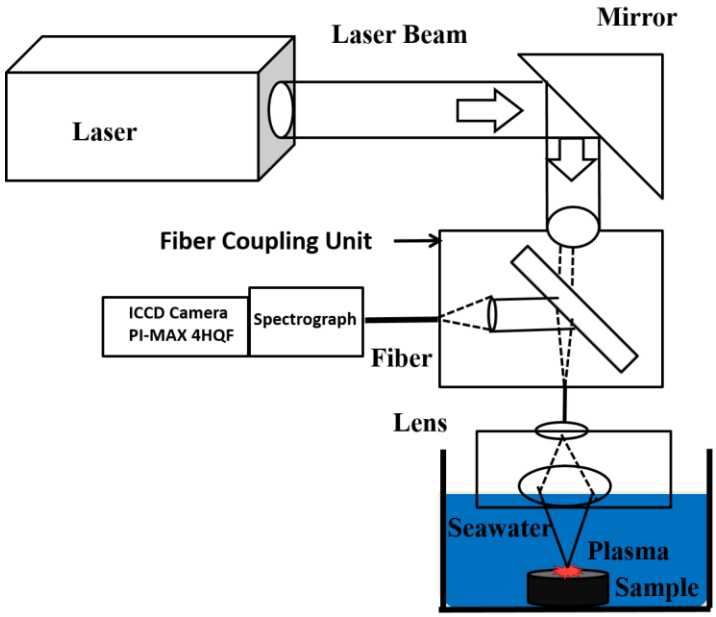


Fig. 3. 1 Experimental Setup

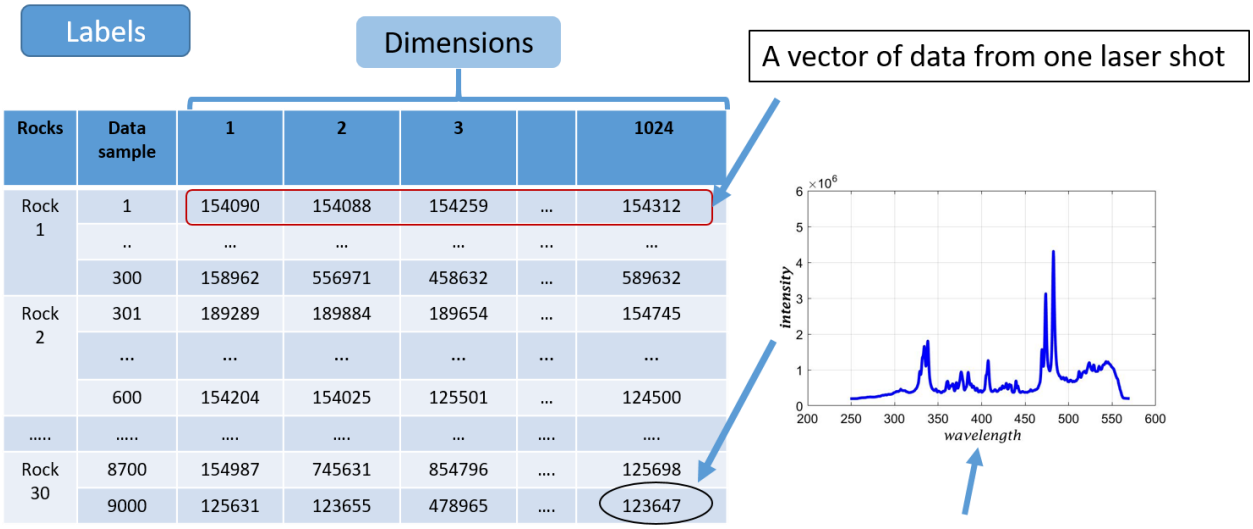


Fig. 3. 2 Matrix of data

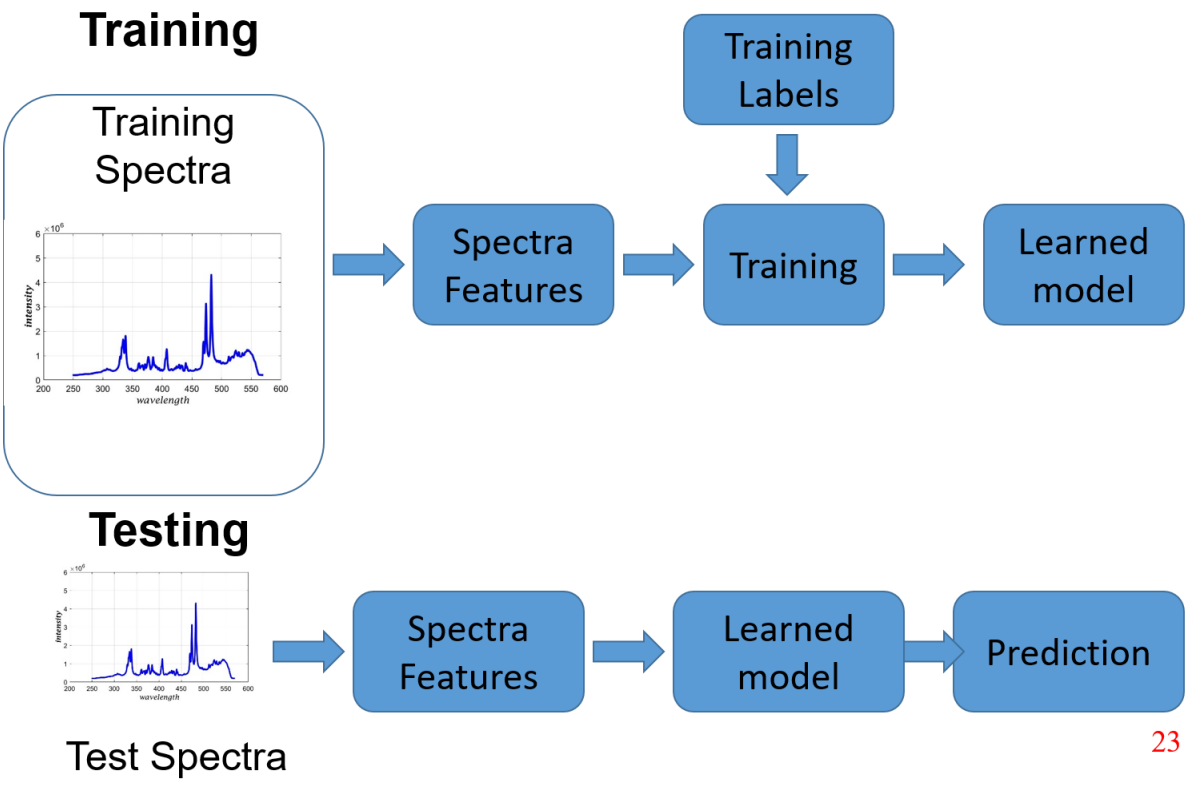


Fig. 3. 3 Model training using machine learning algorithms

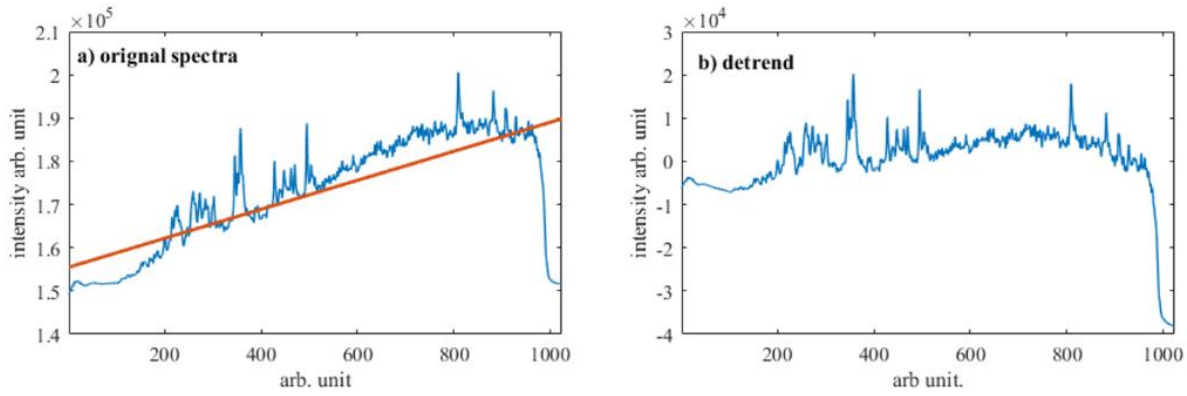


Fig. 3. 4 Effect of detrend operation

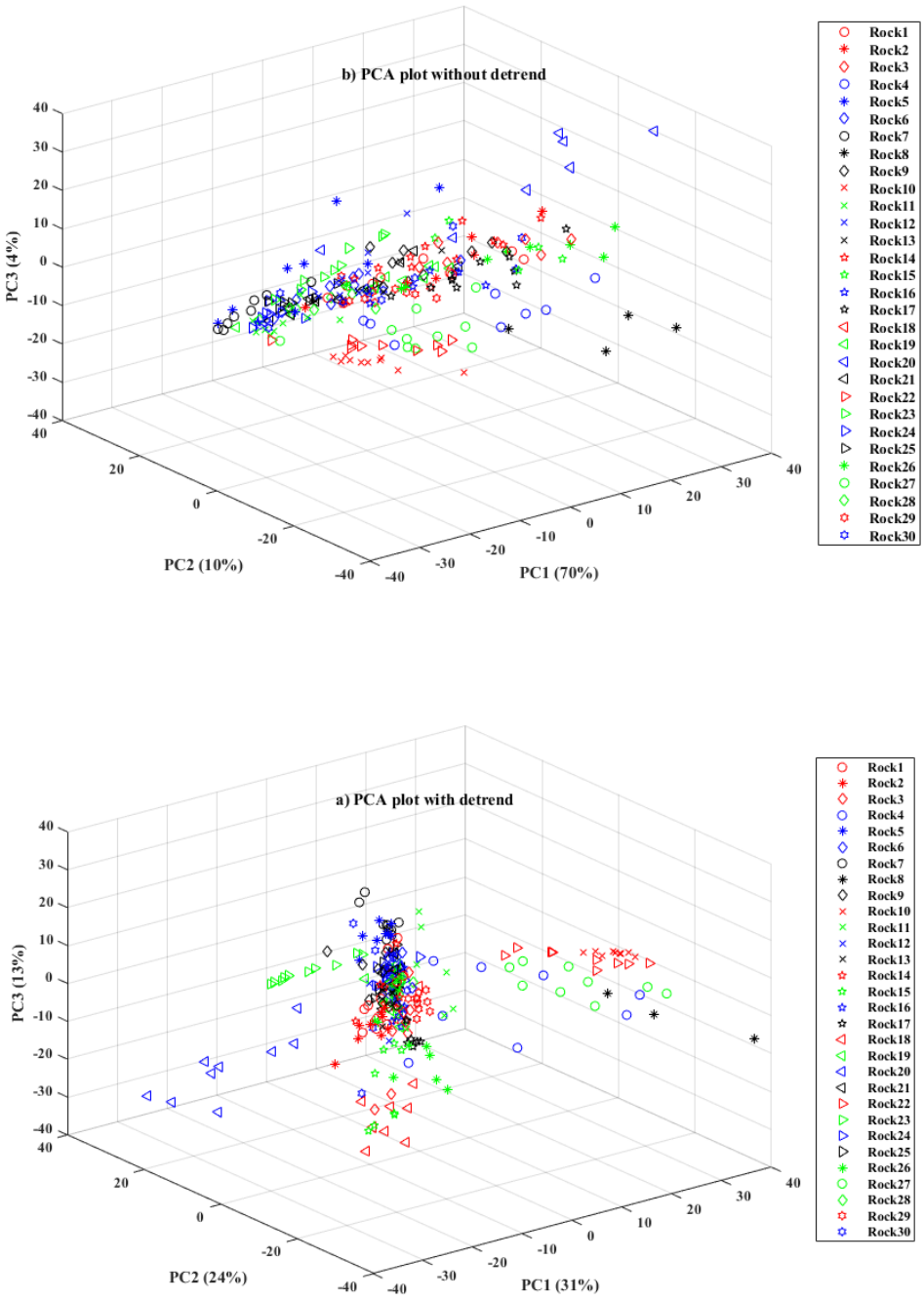


Fig. 3. 5 First three principal component projection of all the rocks (A) without detrend operation and (B) with detrend operation

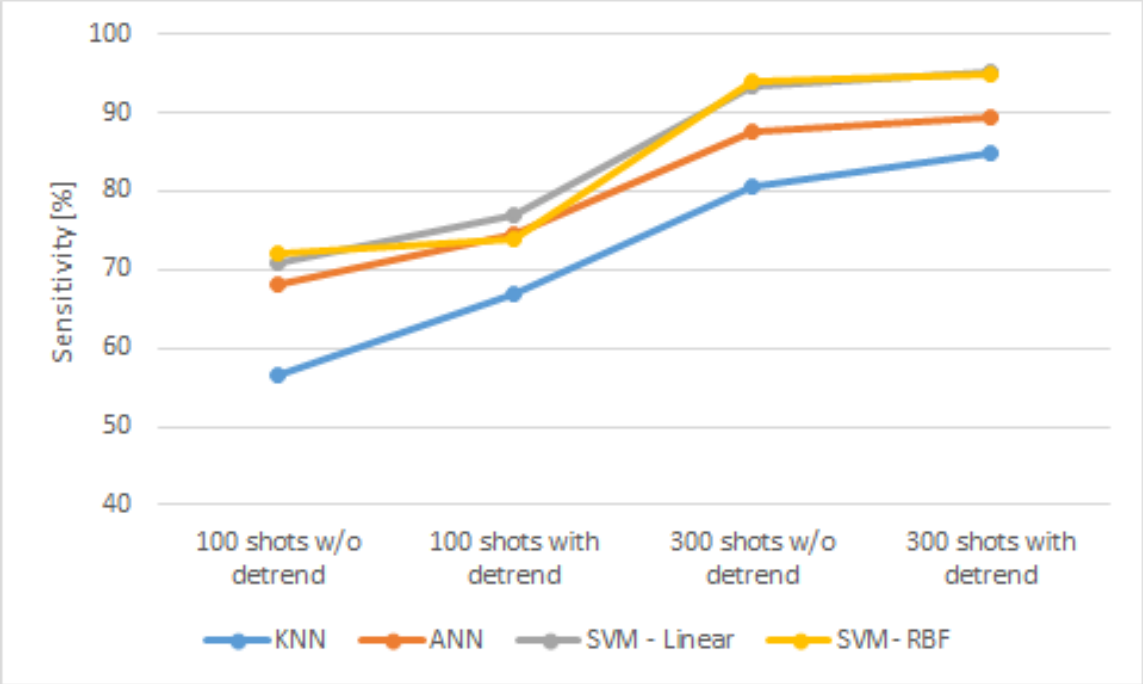


Fig. 3. 6 The sensitivity plot of the four cases for rock label classification

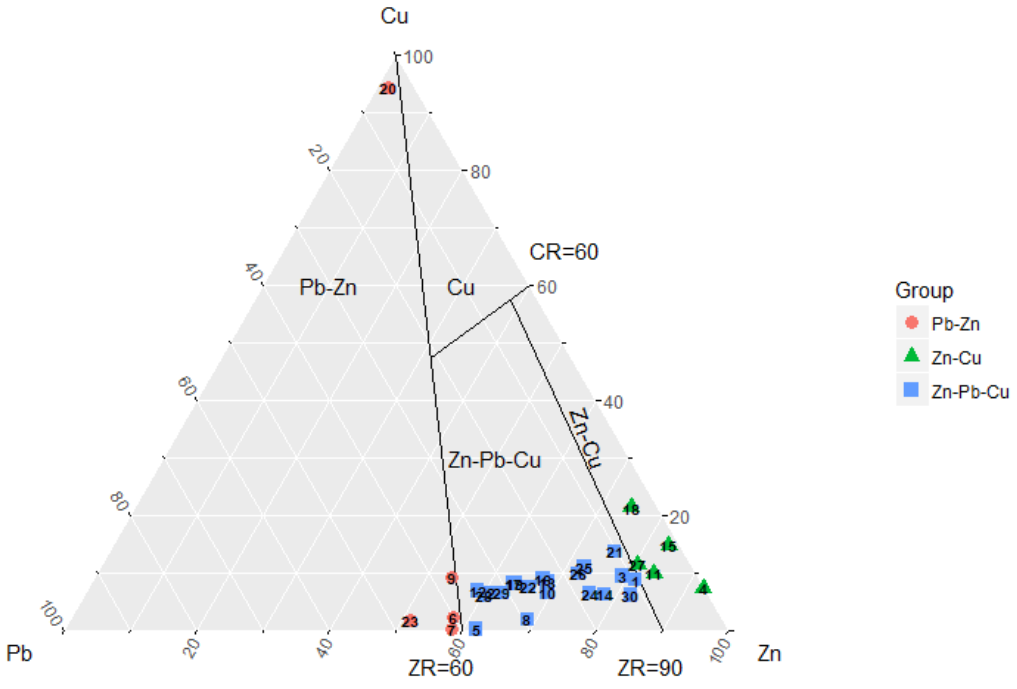


Fig. 3. 7 Ternary diagram plot

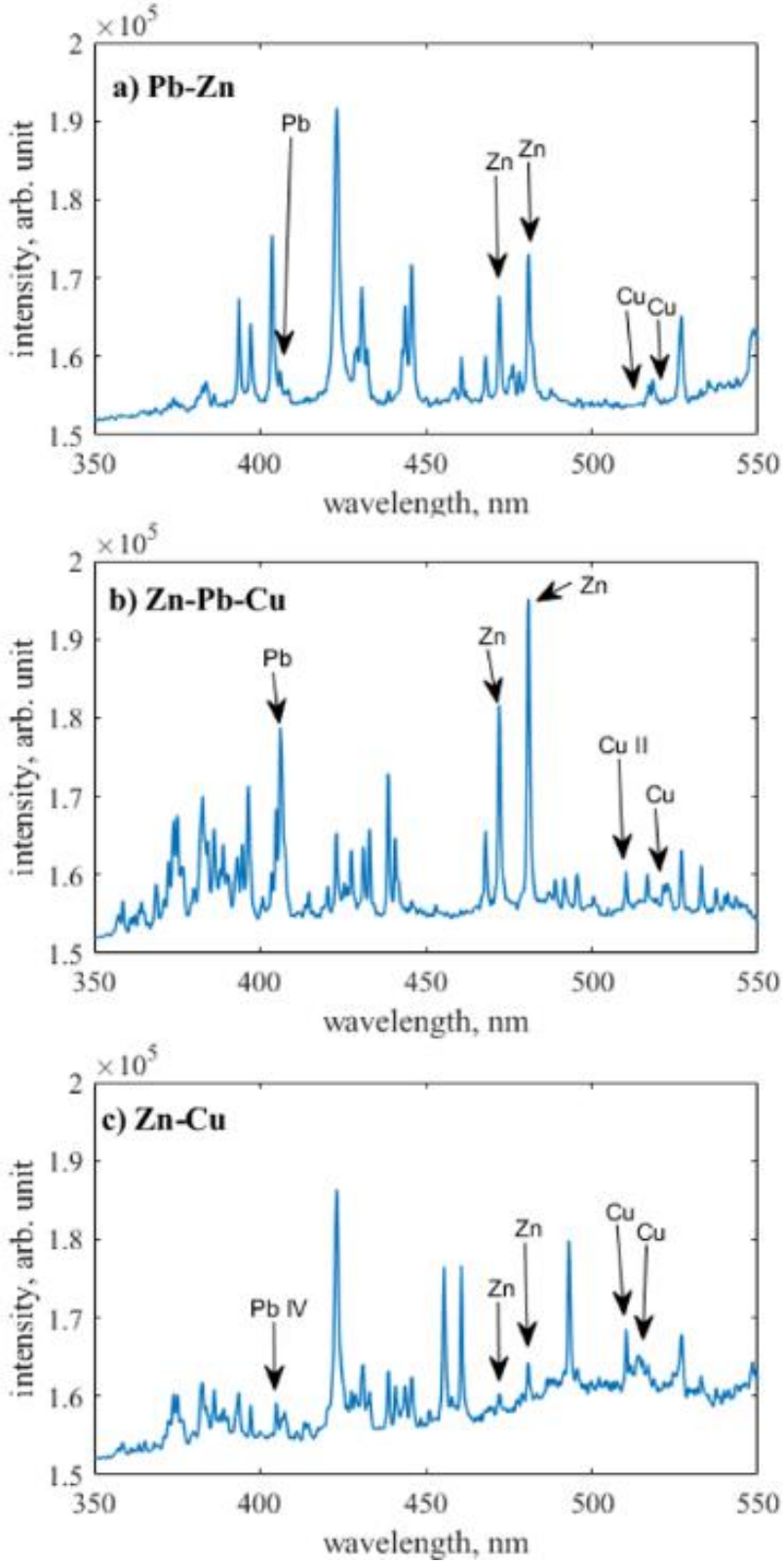


Fig. 3. 8 Spectra of each group rock

Rocks	Spectra	Group	Classification results of label	Classification results of group
6	1st spectra	Pb-Zn	6	Pb-Zn
	2nd spectra	Pb-Zn	20	Pb-Zn
	3rd specra	Pb-Zn	4	Zn-Cu
4	1st spectra	Zn-Cu	4	Zn-Cu
	2nd spectra	Zn-Cu	18	Zn-Cu
	3rd spectra	Zn-Cu	6	Pb-Zn
20	1st spectra	Pb-Zn	20	Pb-Zn
	2nd spectra	Pb-Zn	6	Pb-Zn
18	1st spectra	Zn-Cu	18	Pb-Zn
	2nd spectra	Zn-Cu	4	Zn-Cu

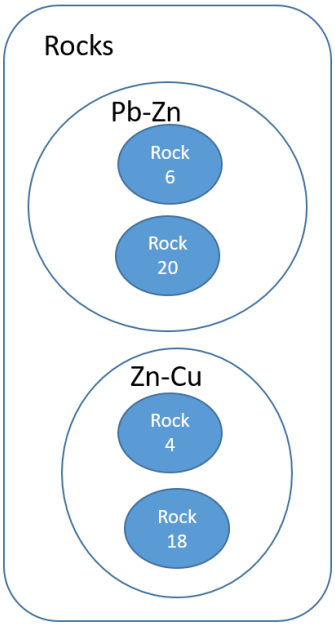


Fig. 3. 9 Rock label vs Rock Group

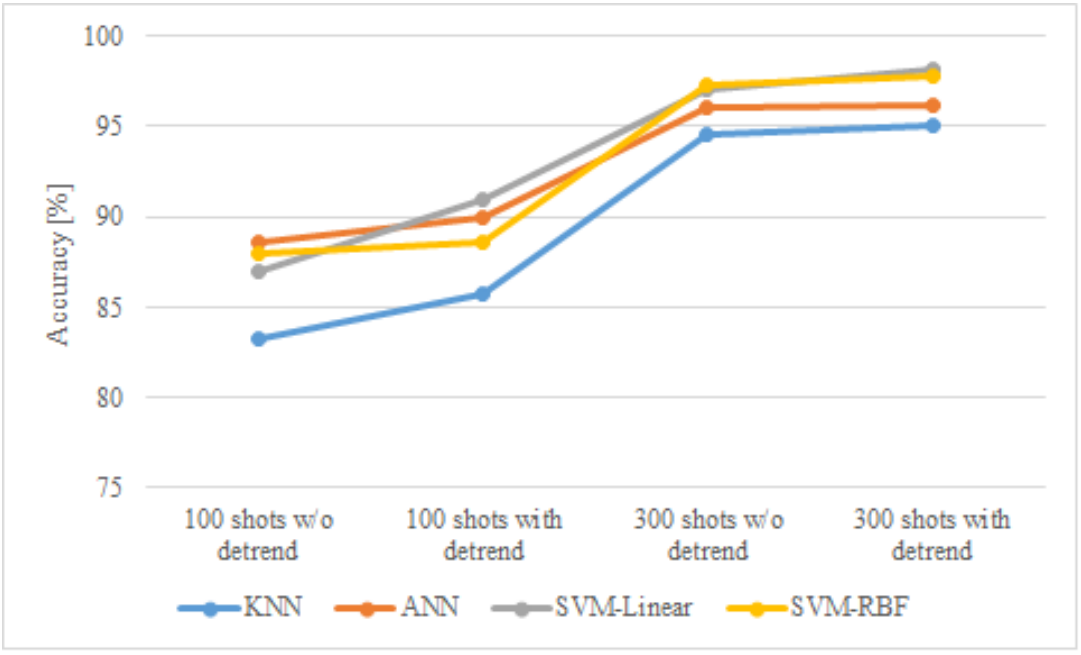


Fig. 3. 10 The average accuracy plot of the four cases for rock group classification

3.8 Tables

Table 3. 1 Variation in the model parameters

KNN	ANN	SVM Linear	SVM-RBF
$N_{PC} \in \{1,2,\ldots 400\}$	$N_{PC} \in \{1,400\}$	$N_{PC} \in \{1,400\}$	$N_{PC} \in \{1,400\}$
$N_{NP} \in \{1,2,\ldots.20\}$	$N_{HL} \in \{4,5,\ldots 100\}$	$C \in \{2^{-5}, 2^{5}\}$	$C \in \{2^{-8}, 2^{20}\}$ $\sigma = \{2^{-20}, 2^{3}\}$

Abbreviations:

N_{PC}: Number of principal components

N_{NP}: Nearest neighbour points

N_{HL}: Number of hidden layers

w/o: without

Table 3. 2 Summary of the best parameters

	KNN	ANN	SVM- Linear	SVM- RBF
100 shots w/o detrend	**N_{PC} = 23** **N_{NP} = 8** **PerInfo=94.2%**	**N_{PC} = 16** **N_{HL} = 80** **PerInfo=93.5%**	**N_{PC} = 336** **$C = 2^{1.8}$** **PerInfo=99.9**	**N_{PC} = 368** **$C = 2^{13}$** **$\sigma = 2^{-18}$** **PerInfo=99.9%**

100 shots with detrend	**N_{PC} = 28** **N_{NP} = 5** **PerInfo=97.8%**	**N_{PC} = 24** **N_{HL} = 82** **PerInfo=97.5%**	**N_{PC} = 388** **C = $2^{-0.6}$** **PerInfo=99.8%**	**N_{PC} = 360** **C = 2^{14}** **$\sigma = 2^{-15}$** **PerInfo=99.8%**
300 shots w/o detrend	**N_{PC} = 10** **N_{NP} = 6** **PerInfo=92.1%**	**N_{PC} = 19** **N_{HL} = 65** **PerInfo=93.8%**	**N_{PC} = 312** **C = $2^{1.1}$** **PerInfo=99.9%**	**N_{PC} = 323** **C = 2^{12}** **$\sigma = 2^{-9}$** **PerInfo=99.9%**
300 shots with detrend	**N_{PC} = 17** **N_{NP} = 3** **PerInfo=96.8%**	**N_{PC} = 15** **N_{HL} = 78** **PerInfo=96.4%**	**N_{PC} = 345** **C = $2^{-2.8}$** **PerInfo=99.8%**	**N_{PC} = 370** **C = 2^{14}** **$\sigma = 2^{-11}$** **PerInfo=99.8%**

Abbreviations:

N_{PC}: Number of principal components

N_{NP}: Nearest neighbor points

N_{HL}: Number of hidden layers

w/o: Without

PerInfo: Percentage of information from original data

Table 3. 3 Summary of the average classification accuracies sensitivities and specifies in percentage (%)

	KNN	ANN	SVM- Linear	SVM- RBF
100 shots w/o detrend	**Sensi = 56.5 Specifi = 99.7 Acc=97.1**	**Sensi = 68 Specifi = 99.7 Acc=97.8**	**Sensi = 71 Specifi = 99.8 Acc=98**	**Sensi = 72 Specifi = 99.8 Acc=98.1**
100 shots with detrend	**Sensi = 67 Specifi = 99.7 Acc=97.8**	**Sensi = 74.6 Specifi = 99.7 Acc=98.3**	**Sensi = 77 Specifi = 99.8 Acc=98.4**	**Sensi = 74 Specifi = 99.8 Acc=98.2**
300 shots w/o.detrend	**Sensi = 80.5 Specifi = 99.3 Acc=98.7**	**Sensi = 87.5 Specifi = 99.6 Acc=99.3**	**Sensi = 93.4 Specifi = 99.7 Acc=99.5**	**Sensi = 94.1 Specifi = 99.7 Acc=99.6**
300 shots with detrend	**Sensi = 84.8 Specifi = 99.4 Acc=98.9**	**Sensi = 89.5 Specifi = 99.5 Acc=99.4**	**Sensi = 95.2 Specifi = 99.1 Acc=99.6**	**Sensi = 94.8 Specifi = 99.8 Acc=99.6**

Abbreviations:

w/o: Without

Sensi: Sensitivity

Specifi: Specificity

Acc: Accuracy

Grp Acc: Group accuracy

Table 3. 4 The p-values of the one-sided Wilcoxon signed rank test for rock label classification

	RBF	KNN	ANN
Linear kernel	**0.47**	**<0.01**	**<0.01**

Table 3. 5 Percentage of Principal component for with and without detrend

	PC1 [%]	PC2 [%]	PC3 [%]
Without detrend	70	10	4
With detrend	31	24	13

Table 3. 6 Mass fractions of Cu, Pb and Zn

Rock number	Cu	Zn	Pb	Others	Group
1	4.63	42.6	4.99	47.78	Zn-Pb-Cu
2	3.08	28.5	15.1	53.32	Zn-Pb-Cu
3	3.69	30.9	4.36	61.05	Zn-Pb-Cu
4	1.85	22.9	0.02	75.22	Zn-Cu
5	0.19	43.4	26.5	29.91	Zn-Pb-Cu
6	1.55	37.5	26.2	34.75	Pb-Zn
7	0.009	2.98	2.12	94.89	Pb-Zn
8	0.153	5.21	2.22	92.41	Zn-Pb-Cu
9	6.12	35.1	24.1	34.68	Pb-Zn
10	0.216	2.25	0.77	96.75	Zn-Pb-Cu
11	2.63	21.9	1.61	73.86	Zn-Cu
12	3.65	30.7	17.9	47.75	Zn-Pb-Cu
13	4.66	37.8	12.7	44.84	Zn-Pb-Cu
14	3.6	44.3	8.79	43.31	Zn-Pb-Cu

15	6.3	35.72	0.65	57.28	Zn-Cu
16	4.57	34.06	11.79	49.57	Zn-Pb-Cu
17	4.29	33.28	14.85	47.56	Zn-Pb-Cu
18	8.02	27.98	1.50	62.48	Zn-Cu
19	3.05	23.59	10.31	63.03	Zn-Pb-Cu
20	3.73	0.0679	0.161	96.04	Pb-Zn
21	5.11	28.2	3.79	62.90	Zn-Pb-Cu
22	3.05	26.4	10.5	60.05	Zn-Pb-Cu
23	0.045	1.24	1.14	97.57	Pb-Zn
24	2.13	24.8	5.77	67.30	Zn-Pb-Cu
25	1.77	11.7	2.61	83.92	Zn-Pb-Cu
26	3.36	24.58	5.98	66.06	Zn-Pb-Cu
27	2.34	16.48	1.62	79.54073	Zn-Cu
28	2.09	20.43	11.40	66.07762	Zn-Pb-Cu
29	2.29	21.71	10.64	65.34545	Zn-Pb-Cu
30	3.86	51.19	7.24	37.69415	Zn-Pb-Cu

Table 3. 7 Spectral lines for each group (nm)

Group	Cu	Pb	Zn
Pb-Zn	521.8; 515.3	406.2	472.2;481.1
Zn-Pb-Cu	521.8; 510.3 (Cu II)	406.2	472.2;481.1
Zn-Cu	515.3;510.6	404.9 (Pb IV)	472.2;481.1

Table 3. 8 Group Classification accuracy

	KNN	ANN	SVM- Linear	SVM- RBF
100 shots w/o detrend	Grp Acc=83.3	Grp Acc=88.6	Grp Acc=87	Grp Acc=88
100 shots with detrend	Grp Acc=85.8	Grp Acc=90	Grp Acc=91	Grp Acc=88.6
300 shots w/o.detrend	Grp Acc=94.6	Grp Acc=96	Grp Acc=97	Grp Acc=97.3
300 shots with detrend	Group acc=95	Grp Acc=96.2	Grp Acc=98.2	Grp Acc=97.8

Table 3. 9 The p-values of the one-sided Wilcoxon signed rank test for rock group classification

	RBF	ANN	KNN
Linear	0.3	< 0.01	< 0.01

4 Conclusion and Future Scope

This study demonstrates the use of machine learning algorithms in the classification of hydrothermal rocks obtained from Okinawa Trough. This study is motivated by the need for energy, and limited resources are available to fulfil future demand. The ocean covers 70% of the earth, and most of the ocean is unexplored. Studies show that it contains many energy resources such as hydrothermal vents. Exploring deep ocean is a challenge, and mining ocean too. Hydrothermal vents are around 3000m below the sea surface. It is so below that even sun rays cannot reach there; in fact, this is one reason to form the hydrothermal vents. The recent rise of artificial intelligence and machine learning can help to solve the major issues in exploring the deep ocean. The laser-induced breakdown spectroscopy is a chemical sensing device. The ChemiCam device which is nothing but the LIBS device developed by the University of Tokyo specifically for deep ocean coupled with the technology of artificial intelligence can automate the process of identifying hydrothermal rocks obtained around hydrothermal vents in the deep ocean. Artificial intelligence method such as machine learning and the laser-induced breakdown spectroscopy which is a chemical measurement device can help in identifying the hydrothermal rocks not only concerning labels but also for the group. A Study compared the three algorithms those were SVM, KNN and ANN with an increasing dataset of 30 rocks. The effect of removing linear trend from each spectrum on the classification is verified, and it was found that the use of the 300 shots per rock dataset with the detrend operation shows an improvement in sensitivity. This study demonstrates that not all the

spectral dimensions from the set of hydrothermal rocks contain information that is relevant to group them. The number of dimensions can be efficiently reduced by applying PCA. This study also showed that with a sufficiently large number of PCs and an appropriate classification algorithm, each spectrum of a rock recorded using LIBS could be classified with an accuracy of > 90% using a 300 shots per rock dataset with the detrend operation. The group of rocks is formed based on the composition ratio of Cu-Pb-Zn using a ternary diagram. The performance of the classifiers according to rock group is investigated, and the results indicate that the SVM algorithm performs well with an accuracy > 95% using the 300 shots per rock dataset with the detrend operation.

As future scope, this study can be extended to create a learning model of regression, which can predict the amount of chemical composition in the rock. It is a very challenging task because rocks are not homogenously composite. The rock label and rock group identification along with the rock composition estimation can help to create the chemical map of the ocean.

Appendix-A: Data Collection using Sea Trials

Observation

It was an investigation of the laser-induced breakdown spectroscopy plasma as a method to perform in-situ multi-element analysis of the composition of the liquids and solid deposits on the seafloor. During this cruise, an in situ sensors called the ChemiCam was deployed, which was developed under the 'Program for the development of fundamental tools for the utilization of marine resources' of the Japanese Ministry of Education. ChemiCam uses a technique known as laser-induced breakdown spectroscopy (LIBS) to perform atomic emission spectroscopy on-site at depths of up to 3000m[65]. During NT13-23, was deployed using ChemiCam in Iheya North Field. On-site chemical analysis of both liquids and the exposed surface of hydrothermal deposits blocking the C0013E artificial vent orifice were performed and well-resolved spectra were obtained. The purpose of this cruise was to enhance the operational efficiency of ChemiCam device. The measurements of liquids are common; the measurement of the solids is considered to be complicated because the focal point of the laser has to be precisely aligned with the surface of the solid. To overcome this, a guide laser has been integrated with ChemiCam, together with a laser-focusing system that measures the intensity of light reflected from targets surface and uses this signal to control a linear stage[65]. One of the major disadvantages of LIBS is that it can only measure the exposed surfaces. In order to overcome this issue, we developed a deep-sea grinder to remove the weathered surface of the deposits, so that

measurements can be made of freshly exposed surfaces that are more representative of whole rock composition[65]. The main objective was to use these tools to enable efficiently, 'sub-surface' measurements of hydrothermal deposits using ChemiCam.

The ChemiCam device as shown in fig. 1.6 is used to perform in-situ multi-element analysis of the composition of liquid and solids at the depths of up to 3000m. The focusing probe was attached to ROV manipulator. The manipulator would be taken to the vicinity of the sample, after which laser will be focussed. The manipulator and the laser were controlled by RS232 communication line on the ROV. The data were monitored in real-time.

A deep –sea grinder was deployed using to remove the weathered surface of hydrothermal deposits. The central hydraulic unit of the ROV supplies pressure that was being controlled in control room. The grinder was operated using ROV manipulator.

Table A- 1 Summary of the sea trial

Sr. No.		**January 2016**
1	Device	ChemiCam
2	Cruise number	NT-16-01
3	Chief Scientist	Blair Thornton (Institute of Industrial Science, The University of Tokyo)
4	Drive number	#1928,1929,1930

5	Location	Ihoya North field, Okinawa trough, Japan
6	Depth [m]	1000
7	Ship	R/V Natsushima
8	Target measured	Natural rocks at 12 points (Cu, Fe, Pb, Zn,Ca,Ba)
9	Achievement	Robust measurement of rocks.

ROV operation

There were two main objectives of the dive

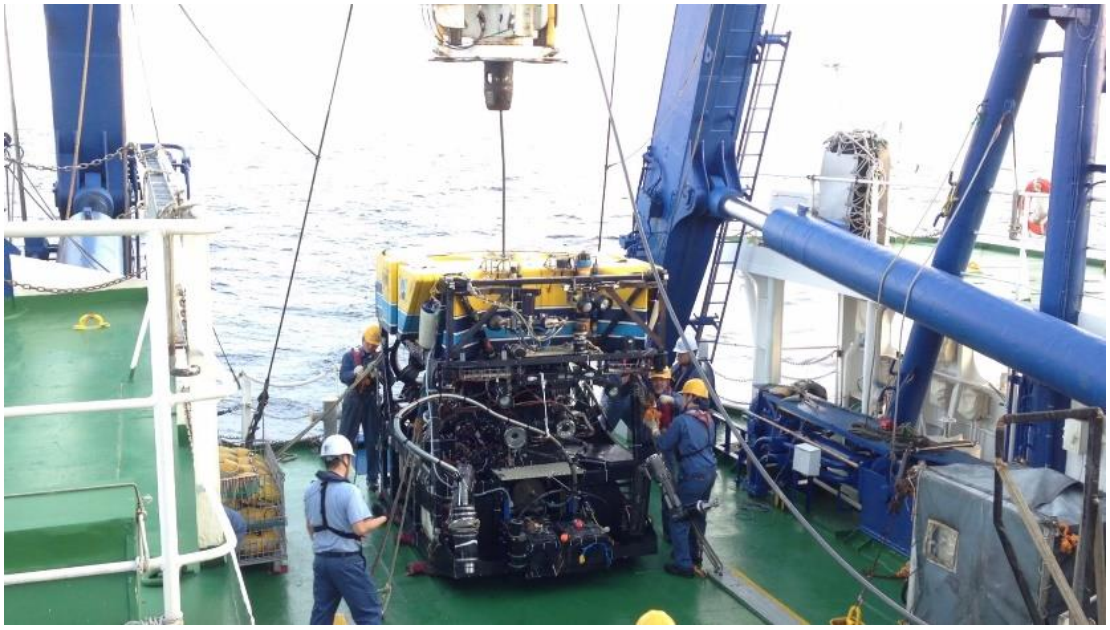
1. ROV dye and
2. To collect the hydrothermal samples from the following locations

    a. Landing point: 27-47, 484N 126-53.821E, 1011m
    b. Landing point: 27-47.505N 126-53.804E, 1080m
    c. Landing point: 27-47.495N 126-53.813E, 1015m

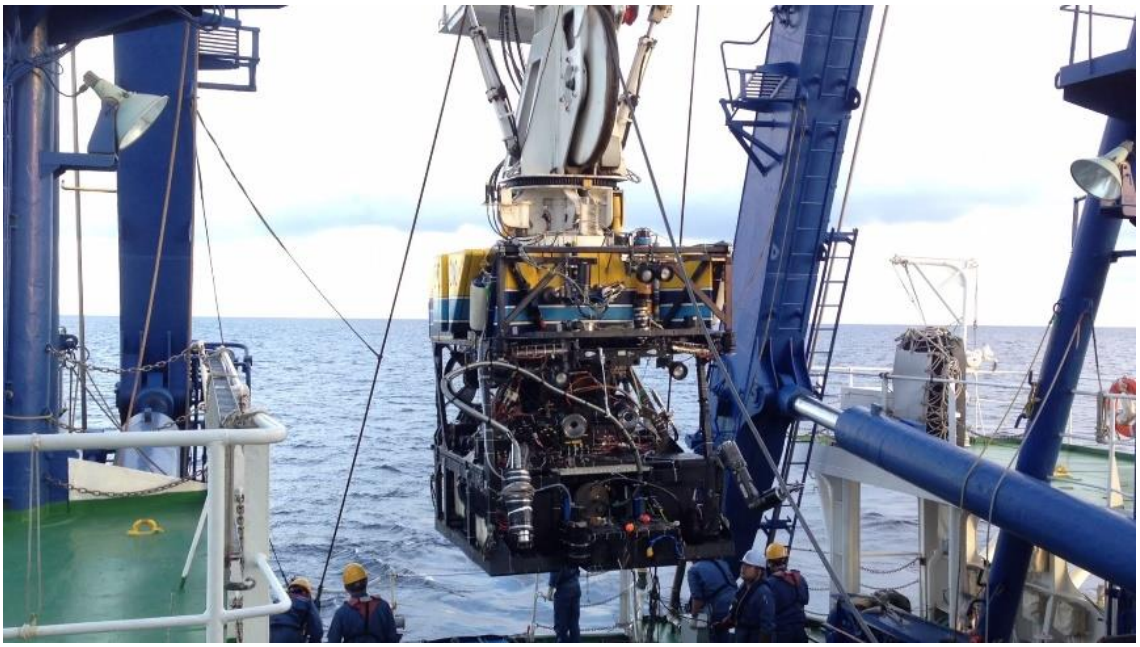
The ROV dolphin was equipped with ChemiCam F device. The ROV manipulator was used to hold the ChemiCam.

The objective of the ROV dye is to collect the hydrothermal samples. The ROV hyper-dolphin (ROV-HPD) is shown in fig. A-1 with ChemiCam device. Fig. A-2 (a) to (d) shows the sequence of the events of the in order that how it is taken into the ocean.

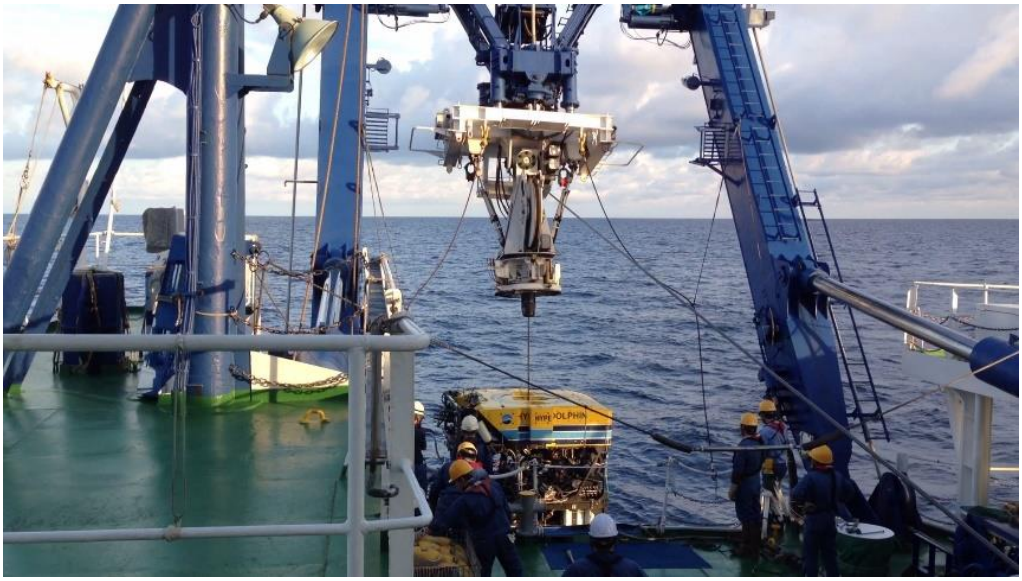
Fig. A-1. 1ROV Hyper-Dolphin (HPD) with ChemiCam device



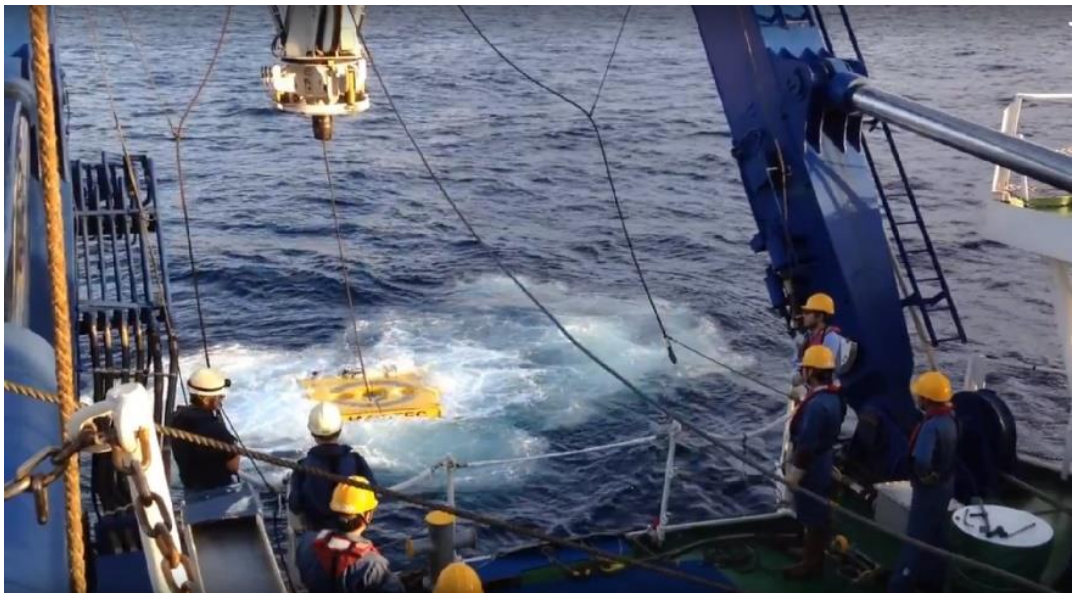
(a)



(b)



(c)

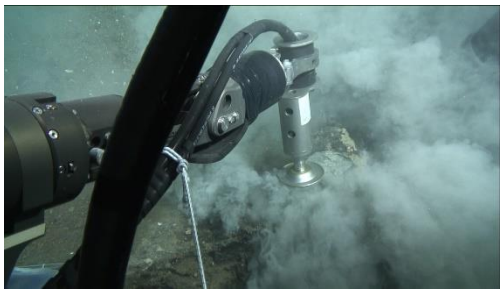


(d)

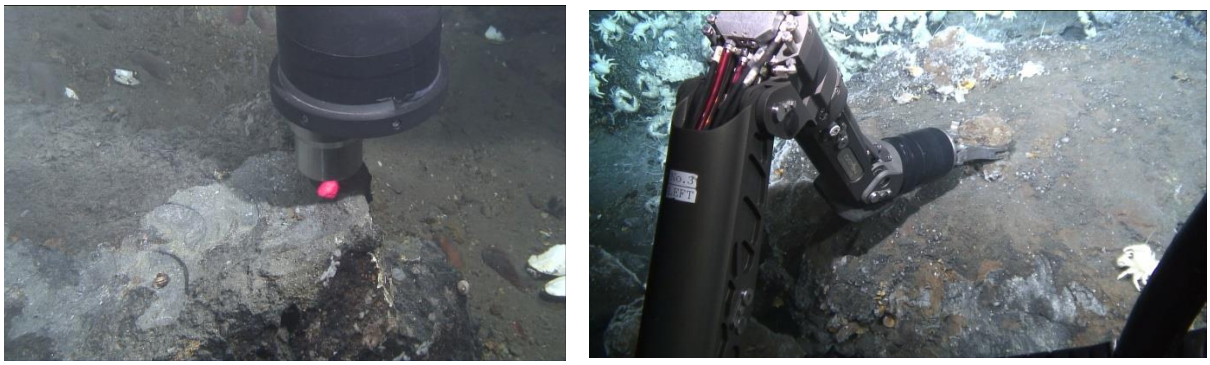
Fig. A-1. 2 (a) to (d) shows the sequence of events of how ROV is controlled by crane for under ocean research

Once the location was decided, grinding is done using grinding machine, attached to the ROV manipulator as shown in fig. A-1

Fig. A-1.



3Grinding machine operating on rocks in the ocean



(a)

(b)

Fig. A-1. 4 (a) and (b) shows the sequence of events in collecting hydrothermal rocks from the deep ocean.

(a)

(b)

Fig. A-1. 5 (a) and (b) shows the sequence of events of how the rocks have been brought back on land and then broken into pieces and packed into a plastic bag for further investigation

Rock Sample list

The sample list of all the operation is shown below. It consists of sample code for each location, the latitude and longitude information has been given from where the rocks have been collected, and the depth in meter has also been shown in the tables A-2

Table A- 2 Samples obtained in the Iheya North field (Operation code 1928)

Sample Code	Latitude	Longitude	Depth (m)
HPD 1928 R1	27 47.476 N	126 53.800E	1003
HPD 1928 R2	27 47.465N	126 53.788E	996
HPD 1928 R3	27 47.465N	126 53.802E	993

Fig. A-1. 6 Samples obtained in the Iheya North Field (#1928)

Table A- 3 Samples obtained in the Iheya North field (Operation code 1929)

Sample Code	Latitude	Longitude	Depth (meter)
HPD 1929 R1	27 47.464N	126 53.823E	1016
HPD 1929 R2	27 47.448N	126 53.796E	999
HPD 1929 R3	27 47.465N	126 53.801E	993
HPD 1929 R4	27 47.465N	126 53.801E	993

Fig. A-1. 7 Samples obtained in the Iheya North Field (#1929)

1) #1930 Operation.

Table A- 4 Samples obtained in the Iheya North Field (#1930)

Sample Code	Latitude	Longitude	Depth (meter)
HPD 1930 R1	27 47.466N	126 53.798E	1001
HPD 1930 R2	27 47.461N	126 53.821E	1013
HPD 1930 R3	27 47.461 N	126 53.821E	1013
HPD 1930 R4	27 47.444 N	126 53.819E	1015
HPD 1930 R5	27 47.430N	126 53.816E	1018
HPD 1930 R6	27 47.452N	126 53.803E	992
HPD 1930 R7	27 47.458N	126 53.800E	1015
HPD 1930 R8	2747.458N	126 53.800E	1015

Fig. A-1. 8 Samples obtained in the Iheya North Field (#1930)

Appendix-B List of Publication

Conference

1. Yelameli M, Thornton B, Takahashi T, Weerkoon T, Takemura Y, Ishii K. Support vector machine based classification of seafloor rock types measured underwater using Laser-Induced Breakdown Spectroscopy. InOCEANS 2016-Shanghai 2016 Apr 10 (pp. 1-4). IEEE.

Journal

1. Yelameli M, Thornton B, Takahashi T, Weerakoon T, Ishii K. Classification and statistical analysis of hydrothermal seafloor rocks measured underwater using laser-induced breakdown spectroscopy. Journal of Chemometrics. 2018:e3092.

References

[1] S. M. Clegg, E. Sklute, M. D. Dyar, J. E. Barefield, and R. C. Wiens, "Multivariate analysis of remote laser-induced breakdown spectroscopy spectra using partial least squares, principal component analysis, and related techniques," *Spectrochim. Acta - Part B At. Spectrosc.*, vol. 64, no. 1, pp. 79–88, 2009.

[2] B. Thornton *et al.*, "Development of a deep-sea laser-induced breakdown spectrometer for in situ multi-element chemical analysis," *Deep. Res. Part I Oceanogr. Res. Pap.*, vol. 95, pp. 20–36, 2015.

[3] R. R. Large, "Australian Massive Sulfide Deposits :," vol. 87, pp. 471–510, 1992.

[4] M. Bollmann *et al.*, "World Ocean Review," *Phys. Rev. E*, vol. 67, p. 232, 2010.

[5] Maribus, "The World Marine Review Marine Resources - Opportunities and Risks," *World Ocean Rev.*, vol. 3, p. 165, 2014.

[6] R. Sydney Avena, Tyler Nowak, *Marine Chemistry: Hydrothermal Vents*. YouTube, 2016.

[7] GRID-Arendal, "Basics of a hydrothermal vent - a Black Smoker," 2014. [Online]. Available: http://www.grida.no/resources/8166. [Accessed: 02-Dec-2018].

[8] GRID-Arendal, "Global distribution of hydrothermal vent fields," 2014. [Online]. Available: http://www.grida.no/resources/8160. [Accessed: 02-Dec-2018].

[9] A. Distèche, "Electrochemical Measurements at High Pressures," *J. Electrochem. Soc.*, vol. 109, no. 11, p. 1084, Nov. 1962.

[10] C. Moore *et al.*, "Optical tools for ocean monitoring and research," 2009.

[11] S. D. Wankel *et al.*, "New constraints on methane fluxes and rates of anaerobic methane oxidation in a Gulf of Mexico brine pool via in situ mass spectrometry," *Deep Sea Res. Part II Top. Stud. Oceanogr.*, vol. 57, no. 21–23, pp. 2022–2029, Nov. 2010.

[12] R. T. Short *et al.*, "FOCUS: FIELD-PORTABLE AND MINIATURE MS Underwater Mass Spectrometers for in situ Chemical Analysis of the Hydrosphere," 2001.

[13] X. ZHang, W. J. Kirkwood, P. M. Walz, E. T. Peltzer, and P. G. Brewer, "A Review of Advances in Deep-Ocean Raman Spectroscopy," *Appl. Spectrosc.*, vol. 66, no. 3, pp. 237–249, Mar. 2012.

[14] S. N. White, R. M. Dunk, E. T. Peltzer, J. J. Freeman, and P. G. Brewer, "In situ Raman analyses of deep-sea hydrothermal and cold seep systems (Gorda Ridge and Hydrate Ridge)," *Geochemistry, Geophys. Geosystems*, vol. 7, no. 5, p. n/a-n/a, May 2006.

[15] T. Takahashi, "Quantitative element analysis of water-submerged solids: Temperature segmented multivariate regression of laser-induced plasma emission," The University of Tokyo, 2017.

[16] "Applied Spectra." [Online]. Available: https://appliedspectra.com/technology/libs.html. [Accessed: 04-Dec-2018].

[17] D. A. Cremers and R. C. Chinni, "Laser-Induced Breakdown Spectroscopy—Capabilities and Limitations," *Appl. Spectrosc. Rev.*, vol. 44, no. 6, pp. 457–506, Nov. 2009.

[18] S. J. Rehse, H. Salimnia, and a. W. Miziolek, "Laser-induced breakdown spectroscopy (LIBS): an overview of recent progress and future potential for biomedical applications," *J. Med. Eng. Technol.*, vol. 36, no. 2, pp. 77–89, 2012.

[19] F. C. De Lucia and J. L. Gottfried, "Rapid analysis of energetic and geo-materials using LIBS," *Materials Today*, vol. 14, no. 6. pp. 274–281, 2011.

[20] J. L. Gottfried, F. C. De Lucia, C. A. Munson, and A. W. Miziolek, "Laser-induced breakdown spectroscopy for detection of explosives residues: A review of recent advances, challenges, and future prospects," *Analytical and Bioanalytical Chemistry*, vol. 395, no. 2. pp. 283–300, 2009.

[21] H. Xia and M. C. M. Bakker, "Reliable classification of moving waste materials with LIBS in concrete recycling," *Talanta*, vol. 120, pp. 239–247, 2014.

[22] E. M. Rodriguez-Celis *et al.*, "Laser induced breakdown spectroscopy as a tool for discrimination of glass for forensic applications," *Anal. Bioanal. Chem.*, vol. 391, no. 5, pp. 1961–1968, 2008.

[23] R. A. Multari, D. A. Cremers, T. Scott, and P. Kendrick, "Detection of pesticides and dioxins in tissue fats and rendering oils using laser-induced breakdown spectroscopy (LIBS)," in *Journal of Agricultural and Food Chemistry*, 2013, vol. 61, no. 10, pp. 2348–2357.

[24] M. Saeki *et al.*, "Development of a fiber-coupled laser-induced breakdown spectroscopy instrument for analysis of underwater debris in a nuclear reactor core," *J. Nucl. Sci. Technol.*, vol. 51, no. 7–8, pp. 930–938, 2014.

[25] J. Rakovský, P. Čermák, O. Musset, and P. Veis, "A review of the development of

portable laser induced breakdown spectroscopy and its applications," *Spectrochim. Acta Part B At. Spectrosc.*, vol. 101, pp. 269–287, Nov. 2014.

[26] S. Kraan, "We are IntechOpen , the world ' s leading publisher of Open Access books Built by scientists , for scientists TOP 1 % Control of a Proportional Hydraulic System," *Intech open*, vol. 2, p. 64, 2012.

[27] H. H. Cho, Y. J. Kim, Y. S. Jo, K. Kitagawa, N. Arai, and Y. I. Lee, "Application of laser-induced breakdown spectrometry for direct determination of trace elements in starch-based flours," *J. Anal. At. Spectrom.*, 2001.

[28] T. Hastie, "The Elements Of Statatical Learning," 2002.

[29] N. L. Lanza *et al.*, "Calibrating the ChemCam laser-induced breakdown spectroscopy instrument for carbonate minerals on Mars," *Appl. Opt.*, vol. 49, no. 13, p. C211, May 2010.

[30] M. Nachon *et al.*, "Calcium sulfate veins characterized by ChemCam/Curiosity at Gale crater, Mars," *J. Geophys. Res. Planets*, vol. 119, no. 9, pp. 1991–2016, Sep. 2014.

[31] M. Darby Dyar, J. M. Tucker, S. Humphries, S. M. Clegg, R. C. Wiens, and M. D. Lane, "Strategies for Mars remote Laser-Induced Breakdown Spectroscopy analysis of sulfur in geological samples," 2011.

[32] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdiscip. Rev. Comput. Stat.*, vol. 2, no. 4, pp. 433–459, Jul. 2010.

[33] D. Pokrajac *et al.*, "Performance of multilayer perceptrons for classification of LIBS protein spectra," in *10th Symposium on Neural Network Applications in Electrical*

Engineering, 2010, pp. 171–174.

[34] T. Vance *et al.*, "Classification of LIBS protein spectra using support vector machines and adaptive local hyperplanes," in *Proceedings of the International Joint Conference on Neural Networks*, 2010.

[35] T. Burr, "Pattern Recognition and Machine Learning Pattern Recognition and Machine Learning . Christopher M. Bishop New York : Springer , 2006 . ISBN 0-38731073-8 . xx + 738 pp. 74.95 .," *J. Am. Stat. Assoc.*, vol. 103, no. 482, pp. 886–887, 2008.

[36] J. Cisewski, E. Snyder, J. Hannig, and L. Oudejans, "Support vector machine classification of suspect powders using laser-induced breakdown spectroscopy (LIBS) spectral data," *J. Chemom.*, vol. 26, no. 5, pp. 143–149, 2012.

[37] J. El Haddad *et al.*, "Artificial neural network for on-site quantitative analysis of soils using laser induced breakdown spectroscopy," *Spectrochim. Acta - Part B At. Spectrosc.*, vol. 79–80, pp. 51–57, 2013.

[38] T. Zhang *et al.*, "Quantitative and classi fi cation analysis of slag samples by laser induced breakdown spectroscopy and partial least square (PLS) methods," *J. Anal. At. Spectrom.*, vol. 30, pp. 368–374, 2015.

[39] R. C. Wiens *et al.*, "The ChemCam Instrument Suite on the Mars Science Laboratory (MSL) Rover: Body Unit and Combined System Tests," *Sp. Sci Rev*, vol. 170, pp. 167–227, 2012.

[40] R. E. Boschen, A. A. Rowden, M. R. Clark, and J. P. A. Gardner, "Mining of deep-sea seafloor massive sulfides: A review of the deposits, their benthic communities,

impacts from mining, regulatory frameworks and management strategies," *Ocean Coast. Manag.*, vol. 84, pp. 54–67, Nov. 2013.

[41] B. Thornton, T. Masamura, T. Takahashi, T. Ura, T. Sakka, and K. Ohki, "A study of laser-induced breakdown spectroscopy for analysis of the composition of solids submerged at oceanic pressures," *Ocean. 2011*, pp. 1–7, 2011.

[42] T. Takahashi, B. Thornton, K. Ohki, T. Sakka, and K. Suzuki, "Investigation of Long-Pulse Laser-Induced Breakdown Spectroscopy for Analysis of the Composition of Rock and Sediment Samples Submerged in Seawater," *Ocean. 2013*, pp. 1–6, 2013.

[43] T. Takahashi, B. Thornton, T. Sato, T. Ohki, K. Ohki, and T. Sakka, "Spectrochimica Acta Part B Temperature based segmentation for spectral data of laser-induced plasmas for quantitative compositional analysis of brass alloys submerged in water &," *Spectrochim. Acta Part B*, vol. 124, pp. 87–93, 2016.

[44] K. Varmuza and P. Filzmoser, *Introduction to multivariate statistical analysis in chemometrics*, vol. 64, no. 4. 2010.

[45] I. T. J. Springer, "Principal Component Analysis, Second Edition."

[46] F. Wood, "Principal Component Analysis," 2009.

[47] N. J. Nilsson, "Mlbook," pp. 1–188, 2010.

[48] D. Conway and J. M. White, *Conway - Machine Learning for Hacker - 2012*. .

[49] "What is Semi-Supervised Learning?" [Online]. Available: https://www.datascience.com/blog/what-is-semi-supervised-learning. [Accessed: 12-Jan-2019].

[50] T. Hastie, R. Tibshirani, and J. Friedman, “The Elements of Statistical Learning,” *Elements*, vol. 1, pp. 337–387, 2009.

[51] X. Wu *et al.*, “Top 10 algorithms in data mining,” *Knowl Inf Syst*, vol. 14, pp. 1–37, 2008.

[52] C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2006.

[53] M. Yelameli, B. Thornton, T. Takahashi, T. Weerakoon, and K. Ishii, “Classification and statistical analysis of hydrothermal seafloor rocks measured underwater using laser-induced breakdown spectroscopy,” *J. Chemom.*, no. December 2017, p. e3092, 2018.

[54] C. M. Bishop, “Neural networks for pattern recognition,” *J. Am. Stat. Assoc.*, vol. 92, p. 482, 1995.

[55] R. Gutierrez-Osuna, “Lecture 13: Validation Motivation The Holdout Re-sampling techniques Three-way data splits Intelligent Sensor Systems.” .

[56] F. Wilcoxon, “Individual Comparisons by Ranking Methods,” 1945.

[57] S. Siegel and N. J. Castellan, “Non-Parametric Statistics for the behavioural Sciences,” *MacGraw Hill Int.* 1988.

[58] F. Wilcoxon, “Individual Comparisons by Ranking Methods Author (s): Frank Wilcoxon Published by : International Biometric Society,” *Biometrics Bull.*, vol. 1, no. 6, pp. 80–83, 1945.

[59] “Wilcoxon signed-rank test.” [Online]. Available: https://en.wikipedia.org/wiki/Wilcoxon_signed-rank_test. [Accessed: 13-Jan-2019].

[60] T. Takahashi, B. Thornton, K. Ohki, and T. Sakka, “Calibration-free analysis of

immersed brass alloys using long-ns-duration pulse laser-induced breakdown spectroscopy with and without correction for nonstoichiometric ablation," *Spectrochim. Acta Part B At. Spectrosc.*, vol. 111, pp. 8–14, 2015.

[61] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. 2000.

[62] X. Wang, X. Fan, Y. Xu, J. Wu, J. Liang, and Y. Zuo, "[Baseline correction method for Raman spectroscopy based on B-spline fitting].," *Guang Pu Xue Yu Guang Pu Fen Xi*, vol. 34, no. 8, pp. 2117–21, Aug. 2014.

[63] C. Chang and C. Lin, "LIBSVM : A Library for Support Vector Machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, pp. 1–39, 2013.

[64] "NIST Atomic Spectra Database." [Online]. Available: https://www.nist.gov/pml/atomic-spectra-database. [Accessed: 16-Jan-2019].

[65] "R/V Natsushima cruise report NT14-21."

Acknowledgements

Foremost, I would like to express my sincere gratitude to my supervisor Professor Kazuo Ishii for his continuous support, guidance and motivation for my PhD studies and research, and sharing time with me whenever it was needed. His guidance helped me in all the time of research, writing papers and writing of this thesis.

My sincere thanks also go to Associate Professor Blair Thornton for sharing his ideas from the beginning of my research study and providing me with valuable feedback all the times.

I want to express thanks to Tomoko Takahashi for guiding operate ChemiCam device, help during cruise sea-trials helping in comments during the research progress discussions.

I am grateful to convey my gratitude to the thesis committee: Prof Doobsub James Jahng, Prof Hiroyuki Miyamoto and Prof Eiji Hayashi. for their encouragement, insightful comments and hard questions.

I want to express my thanks to Associate Professor Shyam Pandey for their help in understanding basic chemistry, spectroscopy concepts. I want to express my thanks to Tharindu Weerakoon for active involvement during project discussion and providing practical tips.

I am grateful to convey my gratitude to my friends, Nishant Koganti, Ravinath Tripathi for being a constant source of motivation, who are like my family in Japan.

I want to express thanks to Takumi Robotics Company management members for allowing me to take short-term leave from the office to prepare for the thesis preparation.

I thank my colleagues of ISHII laboratory for their fruitful discussions and support for my research and experiments at different stages.

Last but not least, I would like to thank my dear parents and brother who were always there cheering and encouraging me up and stood by me through the good times and bad to achieve all success.