DISCRIMINATION OF SWERTIA CHIRAYITA USING NEAR INFRARED SPECTROSCOPY TECHNIQUE

A thesis submitted in partial fulfilment of the requirements for the award of the degree of M.Tech in Instrumentation and Electronics Engineering by

SAWON BHOWMIK

Examination Roll Number: M4IEE21010

Registration No. 150015 of 2019-2020

Class Roll Number: 001911103010

Under the Guidance of

Dr. Bipan Tudu

Department of Instrumentation and Electronics Engineering

Faculty of Engineering and Technology

Jadavpur University

Kolkata 700106

July, 2021

JADAVPUR UNIVERSITY
FACULTY OF ENGINEERING AND TECHNOLOGY

Certificate of Recommendation

*I hereby recommend that the thesis titled "**Discrimination of swertia chirayita using near infrared spectroscopy technique**"carried out under my supervision by Mr. Sawon Bhowmik may be accepted towards partial fulfillment of the requirement for the degree of "Master of Technology in Instrumentation and Electronics Engineering" of Jadavpur University.*

Bipan Tudu

Professor
Dept. of Instrumentation & Electronics Engg
Jadavpur University
Saltlake, 2nd Campus
Kolkata-700 106

....
(Dr. Bipan Tudu)
Professor
Instrumentation and Electronics
Engineering, Jadavpur University

Countersigned by:

Sankar Narayan Patra

Dr Sankar Narayan Patra
Head
Instrumentation and Electronics Engineering Dept.,
Jadavpur University Kolkata -700106

....

Head of The Department
Instrumentation and Electronics
Engineering
Jadavpur University

Bijan Sarkar.

....

Dean
Faculty Council of Engineering and
Technology
Jadavpur University

Certificate of Approval

The foregoing thesis is hereby approved as a creditable study of an engineering subject carried out and presented in a manner satisfactory to warrant its acceptance as a prerequisite to the Degree for which it is submitted. It is understood that by this approval, the undersigned does not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein but approves the thesis only for the purpose for which it is submitted

Bipan Tudu

Examiners

Acknowledgements

While bringing out this thesis to its final form, I came across a number of people whose contributions in various ways helped my field of research and they deserve special thanks. It is a pleasure to convey my gratitude to all of them.

I take this opportunity to express a deep sense of gratitude towards my guide **Prof. Bipan Tudu**, for providing excellent guidance, encouragement and inspiration throughout the project work. Without his invaluable guidance, this work would never have been a successful one. I feel proud to say that I had the opportunity to work with an exceptionally experienced Professor like him.

I would like to convey my sincerest thanks to all the faculty of IEE, Jadavpur University, for their teachings during these two years of Masters degree, which formed a basis for all the research work that have been accomplished.

I am also highly grateful to **Mr. Suman Midya**, for giving me suggestions and support from an early stage of this research and providing me extraordinary experiences throughout the work. I specially acknowledge him for his advice as and when required during this research.

I would also like to thank the scholars, **Mr. Hemanta Naskar**, **Ms. Shreya Nag** and **Ms. Debangana Das** at the same University, who are recently working in this field of research, for their valuable suggestions and helpful discussions and lending me their invaluable knowledge and support. I am greatly indebted to all the people mentioned above.

Sawon Bhowmik

Sawon Bhowmik

Abstract

This study aims to develop a simple and cost effective technique to discriminate a widely used medicinal plant called Swertia chirayita and to quantify the samples collected from different geographical locations. For this purpose, Near Infrared spectroscopy is used in this study. Another objective was to quantify different pre-processing techniques for the NIR spectral data and finding out the best possible pre-processing that maximizes the separability among the collected samples. Out of twenty-five pre-processing only six taken for this study taking their separability index values into consideration. A comparative study is done on those selected pre-processing techniques. Feature transformation is done on the data using principal component analysis. Finally, a functional linear model is calibrated for all the selected pre-processed data by using least square regression technique and then model validation and prediction error calculation is done on the trained model.

Content

List of Figures

List of Tables

CHAPTER ONE: **Introduction and scope of the thesis**

1.1. Introduction

Nature has always played a key role in the lives of people by supplying fresh water and oxygen. Not only that from nature people collected a diversity of valuable forest products for food and medicine. Historically plants have played an important role in medicine. Through observation and experimentation, human beings have learned that plants promote health and well-being. The use of these herbal remedies is not only cost-effective but also safe and almost free from serious side effects. Traditionally the village elders, farmers in a tribal society had enormous knowledge about those year-old techniques of healing.

Generations to generation that knowledge passed on and tremendously enriched the field of medicinal plants. That knowledge is still part of medical practices by the people of various regions of Indian sub-continents as well as several other countries including China middle East, Africa, Egypt, South America, and other developing countries of the world. India and China are two of the largest countries in Asia, which have the richest arrays of registered and relatively well-known medicinal plants[1].

According to the WHO, over 80% of the world's population relies on traditional forms of medicine. Largely different parts of plants meet primary health care needs. On the other hand, in India, the collection and processing of medicinal plants and plant products contribute a major part each year to the national economy. People are directly or indirectly involved with plant-based healing both full and part-time. So the economic implication is also present in India around the medicinal plants. WHO also estimated that the present demand for medicinal plants stands approximately at 14 billion US$. The demand for medicinal plant and plant-based products and raw extracts is growing at a steady rate of 15 to 25 percent per year. Also, it is estimated by WHO that the trade of plant-based products is going to be approximately 1 trillion US$ within 2050. In India alone, this trade is approximately 1 billion US$ per year and it is steadily growing and achieving a new high with every passing year. Still, there is a lot to learn about plants and their qualities. Recent estimates suggest the over 9,000 plants have known medicinal applications in various cultures and countries. In Countries that are developing, people rely heavily on these plants but when it comes to using medicinal plants the developed countries are not behind.

Practitioners from Ayurveda to Chinese medicine and the Japanese Kampo system are the user of these plants for their qualities. Allopathic medicine also owes a tremendous debt to medicinal plants; one in four prescriptions filled in a country like the U.S is either synthesized from or derived from plant materials[2]. Then in homeopathy also there is significant use of plants and their extracts. In the figure 1.1 the overall export values of medicinal plant is given.

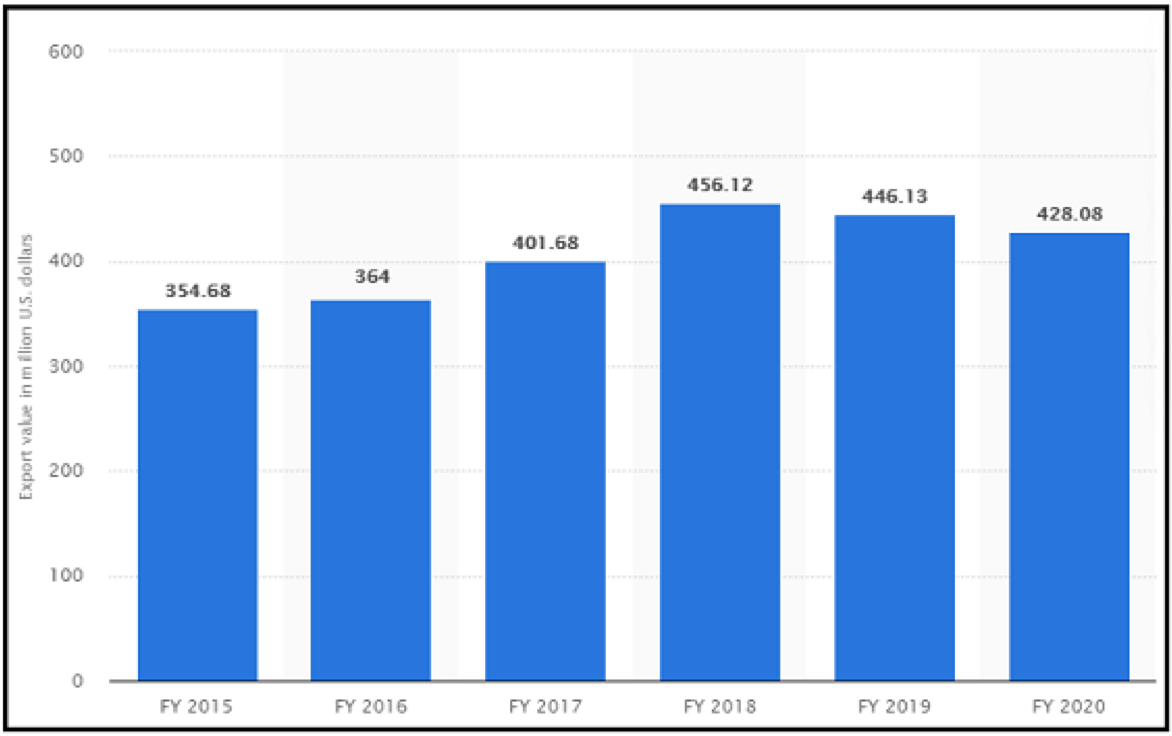
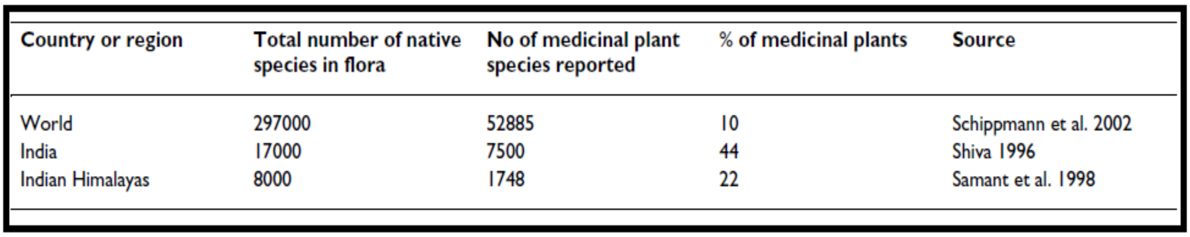


Figure 1.1: year wise export values worldwide bar diagram from 2015 to 2020

Sadly, the research initiative is very limited in determining the quality and benefits of those plants as well as how the plants should be cultivated to get a greater yield that is also unknown for most of the species. Botanical survey of India recorded a total of 15000 plant species in India among which there are approximately 7500 plant species currently in use for their medicinal qualities on a small scale and large scale. Now surprisingly 1700 of those species are recorded or documented for their medicinal properties and drug uses but within that data of only 1200 plants approximately is available[2].

Table 1.1: Country wise reported medicinal plant

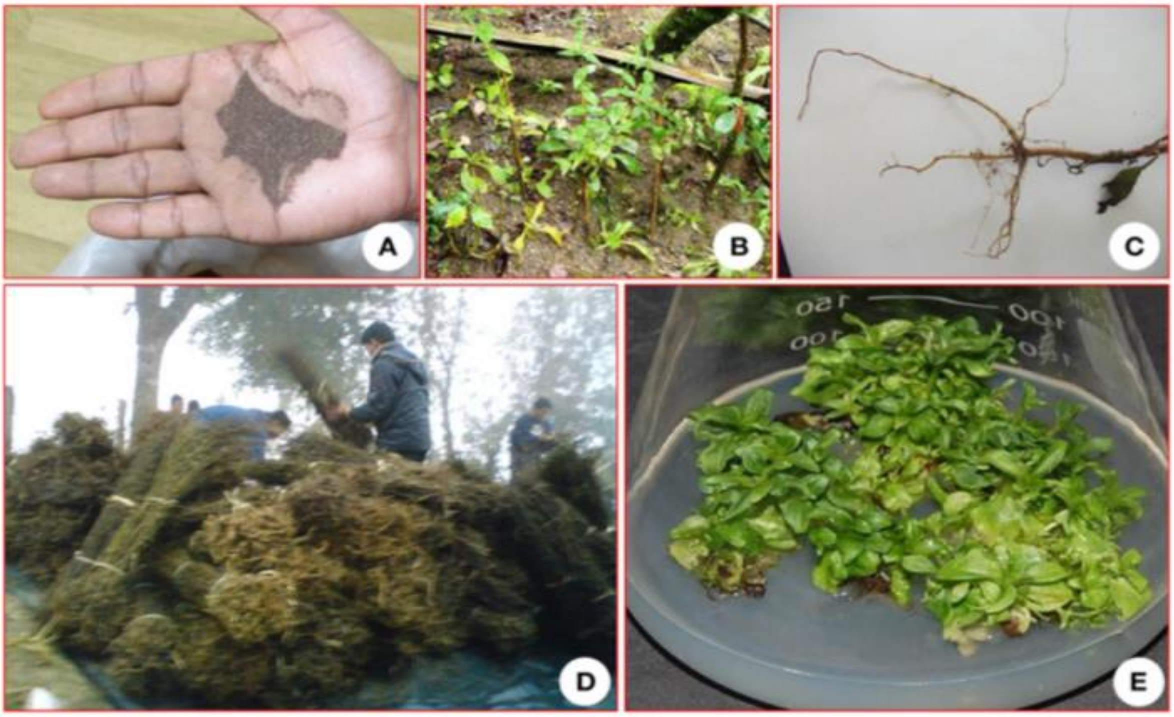


Country or region	Total number of native species in flora	No of medicinal plant species reported	% of medicinal plants	Source
World	297000	52885	10	Schippmann et al. 2002
India	17000	7500	44	Shiva 1996
Indian Himalayas	8000	1748	22	Samant et al. 1998

It can be understood from the table 1.1 that how little is known in this field.

1.2. Introduction to Swertia chirayita

This thesis is based on a particular medicinal plant that has a diverse use in different ailments because of its properties. From the ancient ages, this plant Swertia chirayita is very popular among all practitioners of medicines. Swertia is a genus from the family of Gentianaceae. In India, 40 species of Swertia are recorded. One of those recorded species Swertia chirayita is a well-known and widely used species. Swertia chirayita commonly known as 'Chiretta' is an indigenous plant that grows in the region of the Himalayas at an altitude roughly about 1200 to 2100m mostly on the shady and moist slopes of the mountain. This particular herb is mostly known for its bitter taste. The compound that is mostly responsible for its taste is Amarogentin. The other compounds found are Swerchirin, Swertiamarin[3]. In figure 1.2 different parts of Swertia chirayita is shown.



Swertia chirayita. (A) Seeds, (B) Plant in nature, (C) Root of a mature plant, (D) Dry plant material, (E) High shoot multiplication in a plant tissue culture system.

Figure 1.2: Different parts of Swertia chirayita and high shoot multiplication in plant tissue

For different ailments, the whole plant can be used from its roots to its leaves. Mostly 30 to 40 cm in high the plant has a unique green and purple coloured flower. As discussed earlier that this plant grows in the Himalayas at a particular altitude the states in India where this plant is currently in cultivation are Himachal Pradesh, Uttaranchal, Jammu & Kashmir, Sikkim and Arunachal Pradesh can be seen in the figure 1.3.

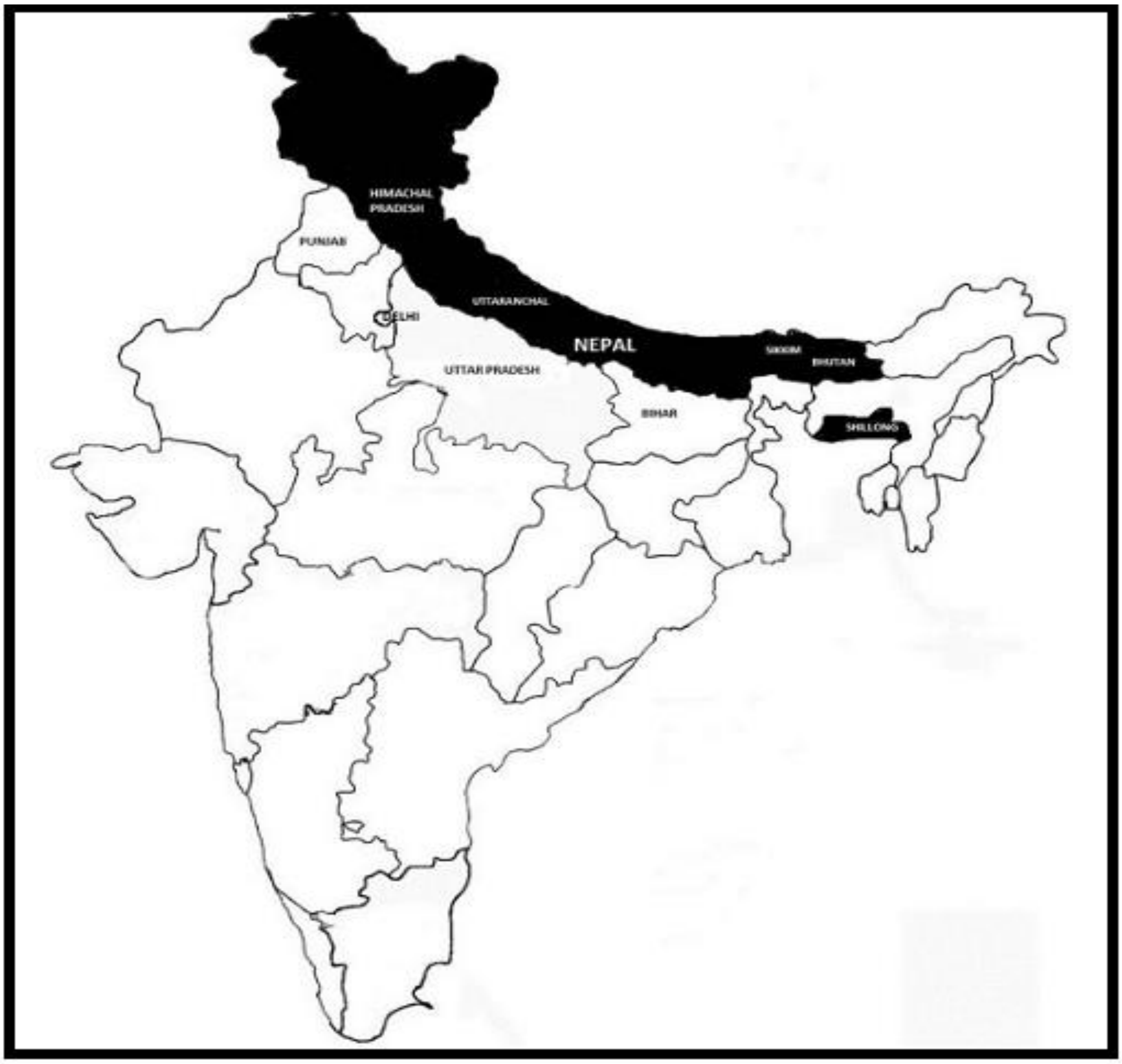


Figure 1.3: regions where Swertia chirayita grows in and around India

1.3. Uses of Swertia Chirayita

Swertia chirayita is a plant widely used in Ayurveda medicines along with that for remedies of different diseases indigenous people in different communities use the whole plant starting from its stem to roots. Mostly used as a treatment of hepatitis, inflammation, and digestive diseases. Its medicinal uses can be found in the treatment of chronic fever, malaria, anaemia, bronchial asthma, hepatotoxic disorders, liver disorders, hepatitis, gastritis, constipation, dyspepsia, skin diseases, worms, epilepsy, ulcers, scanty urine, hypertension, melancholia, and certain types of mental disorders, secretion of bile, blood purification, and diabetes[3,4].

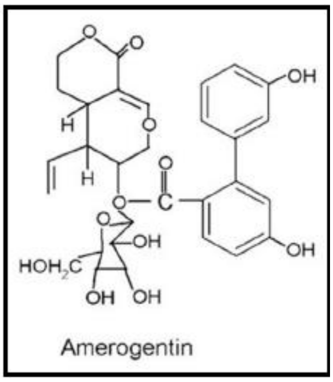
Ethnobotanical uses of Swertia chirayita:

- The whole plant Used in British and American pharmacopoeias as tinctures and infusions.
- The roots are used in treating feature, cough, general weakness etc.
- The roots serve as a treatment of asthma and common cold.
- Leaves and stems if soaked overnight in water. Then made a paste out of it and filtered with one glass of water. If this preparation is taken once in 2 to 3 days this can cure headaches and can be very good for blood pressure[3].
- A paste made out of the herb can be sued for the treatment of skin disease like eczema and pimples.
- For tremor fever the stems and leaves of Swertia chirayita are taken and chopped into small pieces. Boiled with water then depending upon the age of the patient if the required amount of that preparation is consumed it can heal the disease[3].
- Recently it is found that Swertia chirayita have anti-hepatitis qualities along with its anti-fungal and anti-bacterial qualities[3,5].
- The whole plant can be used for the treatment of vomiting, ulcers, gastrointestinal infections, and kidney diseases[3].

1.4. Phytochemistry of Swertia chirayita

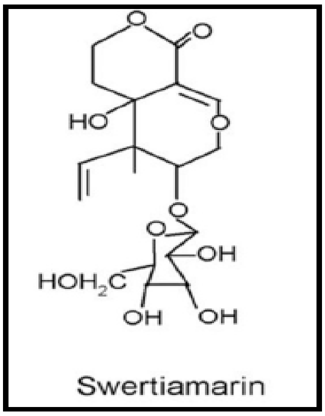
Previously in this discussion we have seen what region does Swertia chirayita grows and also it is discussed that what are the diseases that can be cured in different ways by using this particular herb. Now there is a need of understanding what are the chemical components that makes this plant so unique. There are different phytochemicals that can be found in this plant such as Amarogentin, Swertiamarin, Mangiferin, Swerchirin, Sweroside, Amaroswerin, and Gentiopicrin.

1.4.1. Amarogentin



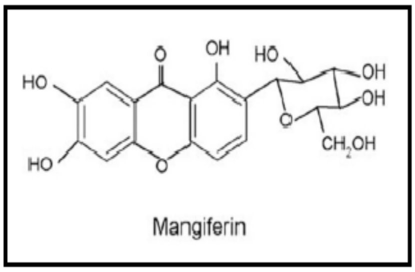
This is the element mostly responsible for the bitter taste of the herb. It is a bitter terpenoid. It has diverse biological activity. This molecule is Antileishmanial, anti-diabetic, gastro-protective. Its molecular formula is $C_{29}H_{30}O_{13}$.

1.4.2. Swertiamarin



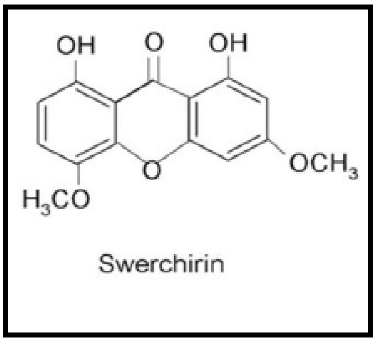
This is another bio active molecule and very effective remedy for different diseases. Swertiamarin is a molecule with anti-cancer, anti-hepatitis, anti-bacterial, anti-diabetic qualities. Its molecular formula is $C_{16}H_{22}O_{10}$.

1.4.3. Mangiferin



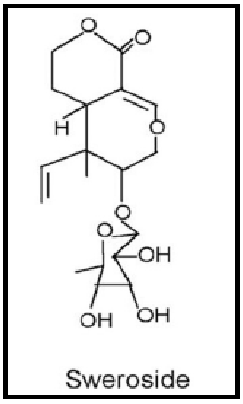
This is the molecule first isolated from leaves and barks of the mango tree. This molecule is a xanthonoid as mangiferin if formed from a xanthone Mangiferin has qualities like antiviral, antioxidant and anti-inflammatory along with that it is also immunomodulatory and anti-diabetic. Its molecular formula is $C_{19}H_{18}O_{11}$.

1.4.4. Swerchirin



Swerchirin also belongs to the class of xanthones. Swertia chirayita and centaurium erythraea are the source of this molecule. It has hypoglycaemic and chemo preventive qualities. Other than that it also proved that it can lower high blood glucose level. Those qualities make it very unique bio active molecule. Its molecular formula is $C_{15}H_{12}O_6$.

1.4.5. Sweroside



Sweroside is another very important bio active compound and it is traditionally used in Chinese medicines for the treatment of osteoporosis. This molecule can be used for its antibacterial qualities along with its hepatoprotective qualities. Molecular formula is $C_{16}H_{22}O_9$.

These are very important bio molecules that are present in swertia chirayita. There are other important molecules like ursolic acid, oleanolic acid, chiratol and Isobellidifolin etc. those molecules also having different medicinal qualities. So overall the plant Swertia chirayita is very important herb[3].

1.5. Challenges

As Swertia chirayita has a huge impact in traditional medicines the demand of this particular herb is growing at a large scale. As well its economic implication is also undeniably huge. This is the reason that the molecule now been over exploited and it is on the verge of extinction.

- Cultivation of this herb is still an issue for the government because this herb can grow at a specific temperature and at a certain altitude. So proper planning in cultivation is not present right now.
- Despite Swertia chirayita is been used traditionally but the safety of its use is still not been established. If there is any harmful toxic element present in the herb is still unknown, so more and more research is necessary in this context.
- As there is a huge demand for Swertia chirayita, there is a rise in adulteration and misidentification as far as this plant is concerned. The use of this plant is wide spread so there must be a measure for quality control but sadly so far there is no such measures in place.
- The presence of previously mentioned molecules is proven but the quantitative measure for those molecules can vary depending upon the place where the plant is cultivated. So a region wise discrimination for the herb is necessary. Depending upon that a quality control can be specified.

1.6. Literature Survey

It is already stated in this introduction that Swertia chirayita is a traditional medicinal plant that has its uses in vast area of medicines. The important aspect of this study is analysis of Swertia chirayita. While a deep study of Swertia chirayita was done during the project, the bio active molecules that are present in chirayita was also studied. As the plant is in the verge of extinction there is a growing need of research on this plant.

- Here the main objective was to explore the potential in medicinal plants resources, to understand the challenges and opportunities with the medicinal plants sector. An elaborate discussion was put in this literature on the economic implications of medicinal plants in the world. there were also suggestions and recommendations based upon the present state of knowledge for the establishment and smooth functioning of the medicinal plants sector along with improving the living standards of the underprivileged communities. This literature reveals that northern India has a diversity of valuable medicinal plants, and attempts are being made at different levels for sustainable utilization of this resource in order to develop the medicinal plants sector[2].

- An overall discussion on Swertia chirayita was stated in this paper. Different ways the stems leave and the roots of the herb can be used for prevention of different ailment was also discussed. The bioactive components present in this plant that was specified and discussed as well how those components are responsible for this plants medicinal qualities were key point in this literature. Also various propositions were made in this paper for planned cultivation and extraction[3].

- Twelve bioactive chemical components were identified and isolated in this literature while studying the chemical components of different extracts of Swertia chirayita. There were different spectroscopy techniques used and their spectra was analysed and the extracts was also mentioned in the literature. Mostly H NMR or photon nuclear magnetic resonance was used other than that IR and C NMR or carbon-13 nuclear magnetic resonance was used for isolating different molecules[6].

- Swertia mussotii and Swertia chirayita are two species of Swertia family they both have uses in traditional Tibetan medicines for treatment of lever and gall bladder disease. Those two were characterized and classified in this literature by using H NMR based metabolomics and well known PCA and PLSR (Partial least square regression technique). A broad range of metabolites, including iridoid glycosides, xanthones, triterpenoids, flavonoids, carbohydrates, and amino acids, were identified and quantified in this study[7].

- This study is based on standardization of the phytochemicals present in Swertia chirayita and upon that distinguish the possibility of adulteration or substitution for the plant. For that in this literature UPLC (Ultra performance liquid chromatography) and PDA (Photodiode array detector) was used, then for chemo-metric data analysis PCA and hierarchical clustering analysis was used for discrimination. Five ecotype of Swertia chirayita was identified and considered as a substitution for chirayita. The data taken from those samples using UPLC were evaluated based on the bio active molecules present in the herb[8].

- This study shows that how NIR spectroscopy is being adopted as a preferred analytical tool to analyse plants due to its minimal sample preparation and shorter analysis time. Also there is a discussion on what factors that can be considered for the analysis and decision making. How NIR eliminates the use of highly hazardous chemical in case of plant analysis. Conventional techniques and its accuracy comparison with the NIRs method was also done in this study[9].

- In this literature how NIRS technique can be used with proper pre-processing and chemo-metrics can be used for pharmaceutical product analysis. all the quantitative qualitative and calibration methods were also discussed in this study. Product specific discussion was done in this literature so to understand when to use PLS, ANN, KNN, MLR, and SVM etc. for different products and plants[10].

- In this study a total of forty samples from three different districts of Nepal was analysed by using Polymerase Chain Reaction (PCR)-based Random amplified polymorphic technique. Genetic diversity between samples were found out by using cluster analysis[11].

1.7. Scope of this research

Swertia chirayita is a widely used medicinal plant for having antioxidant, anti-inflammatory and hepatoprotective abilities in Indian, Chinese and Tibetan medicines. Usually herbal medicines are chopped into small pieces to sell in the market or made powder to use in the medicinal institutes. In that case people confuse this valuable herb with many other herbs. Mixing adulteration is also very common as it is very difficult to identify that using their morphological feature. Sometimes Kalmegh also confused with Swertia chirayita though they have different morphological and phytochemical feature. Also there are different places and altitudes where Swertia chirayita grows that may change its internal bioactive chemical content, so a thorough study should be undertaken to analyse and classify the herb. What are the impact of those adulterations on human health is also unknown. As there is no large scale planning for better cultivation of Swertia chirayita undertaken, this plant is only cultivated by small scale farm owners. Those farmers do not have a simple and well calibrated technique for quantification of Swertia chirayita. As far as Near Infrared Spectroscopy is concerned it is a non-destructive analytical technique. NIRS is known for its simplicity, its rapidity and its low cost. Karl Norris, who had a valuable contribution towards recording the NIR spectra. As well his multivariate treatment of the spectra to determine moisture, oil, protein and starch in biological samples was the first step towards the NIR technology adaption for plant and grain analysis. Another scientist Phil Williams in 1970 had a valuable contribution in NIRS technology, his research on separation of wheat grain according to the protein content is considered to be the first step towards cereal crop analysis using NIR. Now a days NIR technology is widely used for the analysis of plants and plant specific products[12].

The advantages provided by NIR, can be exploited if a proper calibration and classical analytical method is implemented with that. Previously in researches were carried out on this plant using UPLC and PCR techniques for adulteration detection but some those techniques take time to analyse a sample and those techniques are very costly. A simple and cost effective solution is NIR.

1.8. Research discussion:

In this thesis the study of Swertia chirata is undertaken. Samples from six different geographical locations (Nepal, Sikkim) was collected and prepared for NIR analysis. the difference among them was found out from the NIR spectral output. After pre-processing the data, separability index was checked and taking that into consideration the different pre-processing technique’s separability value was compared. Then a classification algorithm was trained with the data and the remaining data was tested and the accuracy of the algorithm was calculated. Then using PCR, a prediction analysis was done on the data collected, and the root mean square error values were calculated.

References:

1. F. Mohammad, “Medicinal Plants of Rural India: A Review of Use by Indian Folks.”

2. C. P. Kala, P. P. Dhyani, and B. S. Sajwan, “Developing the medicinal plants sector in northern India: challenges and opportunities,” *J. Ethnobiol. Ethnomed.*, vol. 2, no. 1, p. 32, Dec. 2006, doi: 10.1186/1746-4269-2-32.

3. V. Kumar and J. Van Staden, “A Review of Swertia chirayita (Gentianaceae) as a Traditional Medicinal Plant.,” *Front. Pharmacol.*, vol. 6, p. 308, 2015, doi: 10.3389/fphar.2015.00308.

4. S. Suryawanshi, N. Mehrotra, R. K. Asthana, and R. C. Gupta, “Liquid chromatography/tandem mass spectrometric study and analysis of xanthone and secoiridoid glycoside composition of Swertia chirata, a potent antidiabetic.,” *Rapid Commun. Mass Spectrom.*, vol. 20, no. 24, pp. 3761–3768, 2006, doi: 10.1002/rcm.2795.

5. L. L. Nyein, “Analysis of different fractions of swertia chirata against gram positive and gram negative bacteria,” *Open Conf. Proc. J.*, vol. 4, no. 1, pp. 29–32, Jan. 2013, doi: 10.2174/2210289201304020029.

6. N. Pant, lain, and Bhakuni, “Some chemical constituents of Swertia.”

7. G. Fan *et al.*, “Metabolic discrimination of Swertia mussotii and Swertia chirayita known as ‘Zangyinchen’ in traditional Tibetan medicine by (1)H NMR-based metabolomics.,” *J. Pharm. Biomed. Anal.*, vol. 98, pp. 364–370, Sep. 2014, doi: 10.1016/j.jpba.2014.06.014.

8. M. Singh *et al.*, “Ultra performance liquid chromatography coupled with principal component and cluster analysis of Swertia chirayita for adulteration check.,” *J. Pharm. Biomed. Anal.*, vol. 164, pp. 302–308, Feb. 2019, doi: 10.1016/j.jpba.2018.10.054.

9. Batten G. D. (1998) Plant analysis using near infrared reflectance spectroscopy: the potential and the limitations. *Australian Journal of Experimental Agriculture* **38**, 697-706.

10. Y. Roggo, P. Chalus, L. Maurer, C. Lema-Martinez, A. Edmond, and N. Jent, “A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies.,” *J. Pharm. Biomed. Anal.*, vol. 44, no. 3, pp. 683–700, Jul. 2007, doi: 10.1016/j.jpba.2007.03.023.

11. J. K. C. Shrestha *et al.*, “Assessment of Genetic Diversity in Nepalese Populations of *Swertia chirayita* (Roxb. Ex Fleming) H. Karst Using RAPD-PCR Technique,” *AJPS*, vol. 04, no. 08, pp. 1617–1628, 2013, doi: 10.4236/ajps.2013.48196.

12. P. K. Sahaya Rajesh, C. Kumaravelu, A. Gopal, and S. Suganthi, “Studies on identification of medicinal plant variety based on NIR spectroscopy using plant leaves,” in *2013 15th International Conference on Advanced Computing Technologies (ICACT)*, Sep. 2013, pp. 1–4, doi: 10.1109/ICACT.2013.6710535.

CHAPTER TWO: **Near Infrared Spectroscopy**

2.1 Introduction to NIR

Near infrared spectroscopy is a spectroscopic method that uses the near infrared region of the electromagnetic spectrum which is from 780nm to 2500nm. This system provides rapid non-invasive and non-destructive and simple analytical data containing information that can be further used for quantification. Application of NIRS can be found in various field from human health to environmental chemistry and plant based product analysis.

2.2. Theory and history of Near Infrared Spectroscopy

Near Infrared Spectroscopy is a kind of vibration spectroscopy. NIR usually employs photon energy more specifically electromagnetic radiation having energy ranges between 2.65 x 10-19 to 7.96 x 10-20 J, the corresponding wavelength range is 750 to 2,500 nm (wavenumbers: 13,300 to 4,000 cm-1). The main objective for the study of Near Infrared Spectrometry is to target those molecules which are showing molecular vibrations. The energy level specified for near infrared radiation is enough for changing the vibration states of those molecules and from that valuable insights as well as a good quantitative analysis can be generated[1,2].

Now going into the history of Near Infrared radiation one needs to go to the discovery of Sir Frederick William Herschel. Sir Herschel was doing an experiment on the contribution of different wavelength lights contained in white sunlight to increase the temperature of the material exposed to the sun. While doing the experiment Sir Herschel did an unusual thing, he didn’t stop experimenting when he reached the visible red colour region he kept going and recorded the temperature with a thermometer, Surprisingly, Temperature kept rising. Herschel used blackened bulb thermometers and glass prisms which are transparent to short wave NIR radiation and reported his achievement by referring to this region as “calorific rays” found beyond the red. Later this ray was named ‘Near Infrared’. In 1950s one can find the application of Near Infrared Spectrometry but it was used with other spectrometric techniques like UV, Mid Infrared etc. In the late 1980s and 90s saw the use of NIRS as a standalone device in the field of chemical analysis and it became more prominent later on with the development of optical fiber[1].

In the early 1930s the NIR Spectroscopy was used to determine the water content in gelatine later Barchewitz was the first one to employ NIR in determination of fuel. In the early days it was expected that beer’s law was obeyed. While Beer’s law states that absorbance is proportional to the path length ‘l’ that the radiation travels through the sample and the concentration of the absorbing species ‘c’[1].

So according to that if beer’s law is written as a function of λ then the function will be

$$A(\lambda) = \epsilon(\lambda) * l * c$$

In later date Karl Norris did a path breaking work using NIR. He was searching new methods for determination of moisture in Agricultural produces. His way of research paved the way for diffused reflectance to be used as a standard for agricultural product analysis and that in turn opens the way of working with the samples directly without any pre-treatment. As well his experiment abandons Beer’s as a prerequisite for quantitative analysis as it was simply not applicable in highly scattering medium[1].

In order to access the origin of a NIR spectra, to be able to interpret it and have an important tool to guide in analytical method development, one should be familiar with the fundamentals of vibrational spectroscopy. The NIR spectrum originates from radiation energy transferred to mechanical energy associated with the motion of atoms held together by chemical bonds in a molecule. Although many would approach method development in a purely empirical way, knowledge of the theory can help to look at the important wavelengths and quicker optimisation of the modelling stage.

Before going deep into vibrational Spectrometry there is a need for discussion of Near Infrared radiation and electromagnetic wave propagation. Maxwell’s theory states that accelerated charges radiate electromagnetic wave. So if we consider a charge oscillating in space with a particular frequency then this produces an oscillating electric field in space, which in turn produces an oscillating magnetic field and this again produces an oscillating electric field. This electric and magnetic field regenerates each other as the wave propagates through the space[4]. The frequency of electromagnetic wave is equal to the frequency of oscillation of the charge.

Classically the electric field and the magnetic field produced are orthogonal to each other here in the below figure a linearly polarized sinusoidal electromagnetic wave is propagating in Z direction and the electric field E is in x direction and the orthogonal magnetic field B is on the y direction.

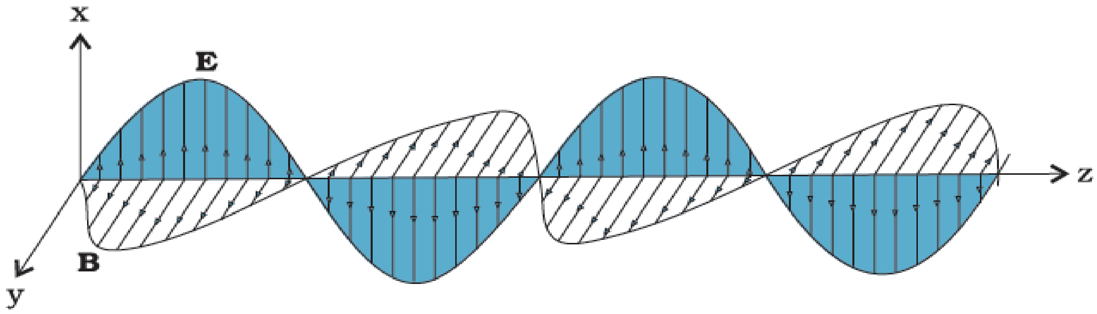


Figure 2.1: Electromagnetic wave propagation.

Now we need to discuss about the spectrum of Electromagnetic radiation, though there is no hard core difference among the different EM waves we can classify them according to their wavelength and frequency.

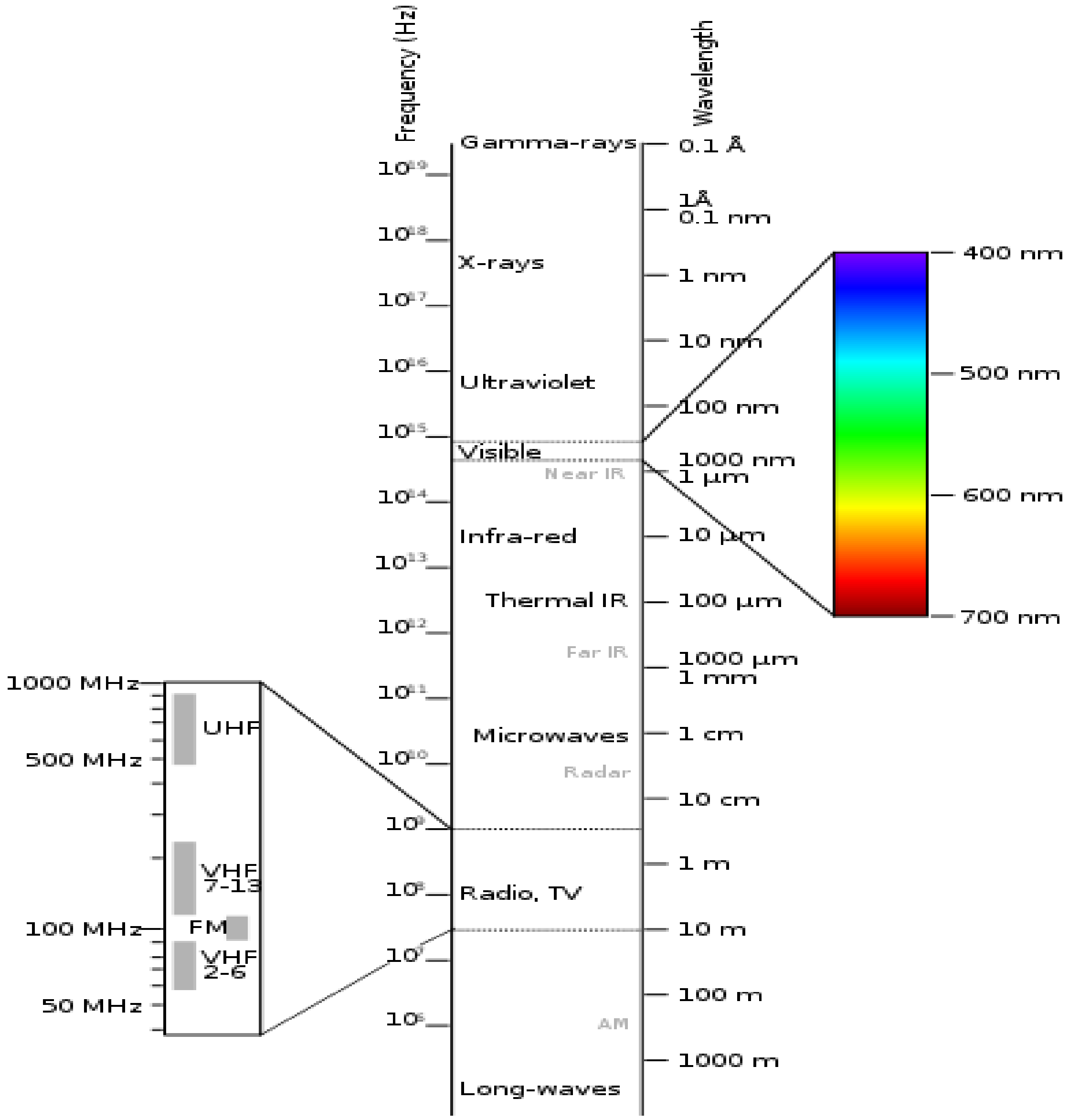


Figure 2.2: Electromagnetic wave spectrum.

We can see that in the figure 2.2 if we come top to bottom the wave length (λ) increasing and the frequency(ν) is decreasing and the spectrum near infrared is shown in the below the visible light spectrum after that starts the infrared region.

2.3. Vibration Spectrometry

As discussed above that near infrared spectrometry can only be employed for a sample that have molecules that show vibrations in their covalent bond structure. The covalent bond in a molecule is not rigid sticks, it is more like a spring. In room temperature or in ambient temperature this bonds may show stretching or rotating motions. Here we are more concerned about the stretching motion in a covalent bond. There are two types of vibrations one is coupled vibration and another is uncoupled vibration. If two molecules in covalent bond and both of the molecule's relative position is changing due to vibrating motion in the bond, then the molecule is in coupled vibration and on the other hand if one molecule is stable in its position and the other molecule is changing its relative position due to vibrating motion then it is in decoupled vibration. Now let's look in to the figure 2.3 for understanding of vibration in a molecule and some concepts on dipole moment[2].

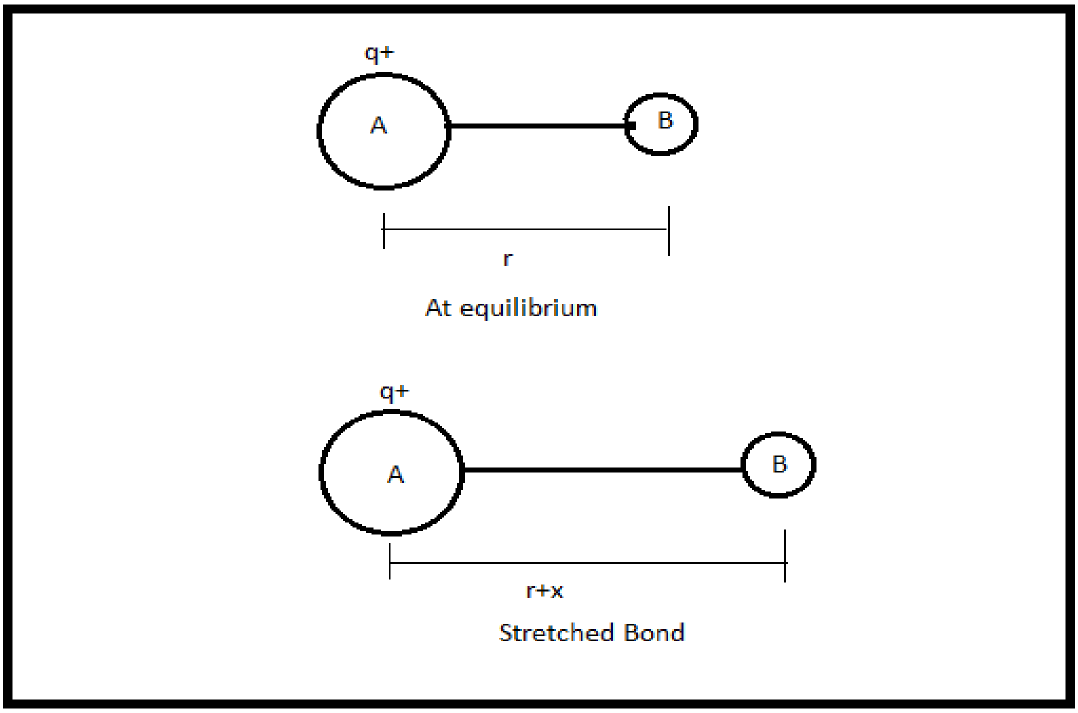


Figure 2.3: vibration within the covalent bond of two molecules

Here suppose A and B are making a covalent bond and they are in vibration and suppose at equilibrium the internuclear distance is r and there is a partial positive and negative charge generated respectively on A and B.

Now dipole moment is a vector quantity which has a magnitude equals to the product of the charge generated (q) and the distance between the two nuclei (r) and the direction is towards the positive to the negative charge.

In the bottom image their bond is been stretched and the distance becomes (r+x) between the nuclei. So it can be seen that the dipole moment also changes. At equilibrium the molecule usually stays at a lower vibration level, later if there is any electromagnetic wave that strikes the molecule having the frequency equal to the frequency of vibration at that instance the molecule can take the energy from the electromagnetic wave.

Atoms which participate in chemical bonds are displaced a certain distance from one another depending upon their bond straight and their individual mass. The amplitudes of these vibrations are of a few nanometres. Now if an electromagnetic radiation of frequency (ν) and energy (E) strikes the molecule then that energy can be transferred to the molecule. The energy (E) can be written as

$E = h * \nu = (h * c)/\lambda$, where λ is the wavelength
h is plank's constant
c is the velocity of light.

2.4. Classical Mechanics

Now in classical mechanics two atoms making a bond is just considered as a spring like arrangement. Here hook's law comes into the picture. Hook's law states that if there is a displacement occurs in an atom in a bond there will be a restoring force generated in the opposite direction to the displacement. Also the restoring force generated will be proportional to the displacement.
So taking that into consideration that force(f) is going to be equal to K * x if x is the displacement. Now for that an energy is going to be produced according to hook's law energy (E) will be

$$E = (h\ /\ 2\pi) * \sqrt{k/\mu}$$

In the equation k is the force constant and h is plank's constant now if we come to μ it is reduced mass. Actually in classical mechanics two atoms are considered as two separate masses assuming one mass is m1 and the other mass is equal to m2 connected together with a string. Then for the whole system the reduced mass will be, μ = (m1*m2)/(m1+m2)[1].

now when we talk about the interaction of molecule and electromagnetic radiation. A molecule in vibration cannot take any arbitrary amount of energy. The interaction and energy transfer depends upon selection rule and behaviour of vibration in a bond and their frequency of vibration. So depending upon the bond structure some of the electromagnetic radiation with a particular frequency and wavelength may interact with a molecule where other electromagnetic wave containing different wavelength may not interact with the molecule that is the theory NIR spectrometry employs to determine the molecule.

2.5. Harmonic and anharmonic oscillator

This vibrating system can be described successfully by ideal harmonic oscillator though more accurate modelling is anharmonic oscillator. Assuming that if two atoms making a bond is in oscillation then periodically their distance changes at a certain frequency. If we consider it is harmonic oscillation, then we are assuming that the distance can change any amount without bond dissociation practically this is not possible.

In this diagram in the x direction we are plotting the interatomic distance and in the y axis we are considering the energy of vibration which is proportional to the frequency of vibration. Now v is called the vibration quantum number.
Form quantum mechanics the formula for the energy(E_{vib}) of the vibration is

$$E_{vib} = (v + ½) * h * ν$$

now in the equation if we put v = 0,1,2,3 then we will get the values for each energy level.
For v = 0 , energy will be equal to 1/2*h*ν
For v = 1 , energy will be equal to 3/2*h*ν
For v = 2 and 3 the energy value respectively is 5/2*h*ν and 7/2*h*ν

So it can be stated that each vibration level has same difference of energy and if an electromagnetic radiation of energy ½*h*ν interacts with the molecule then that molecule can take energy from the radiation and move to the higher vibration level shown in figure 2.4. Now here comes a restriction in the model

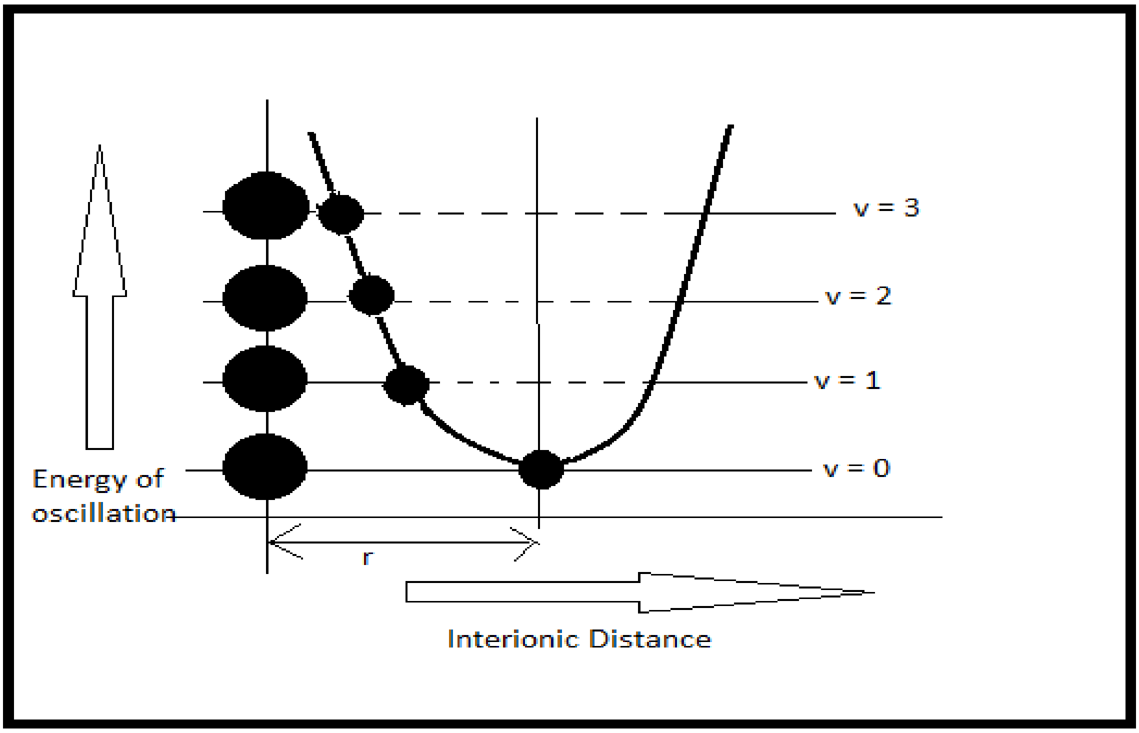


Figure 2.4: harmonic oscillation.

as the molecule can go from one vibration level to its adjacent vibration level this is called the selection rule and according to that Δv will be equal to either +1 or -1. Now in anharmonic oscillator the molecule in covalent bond gets dissociated after a certain energy and the atoms become free.

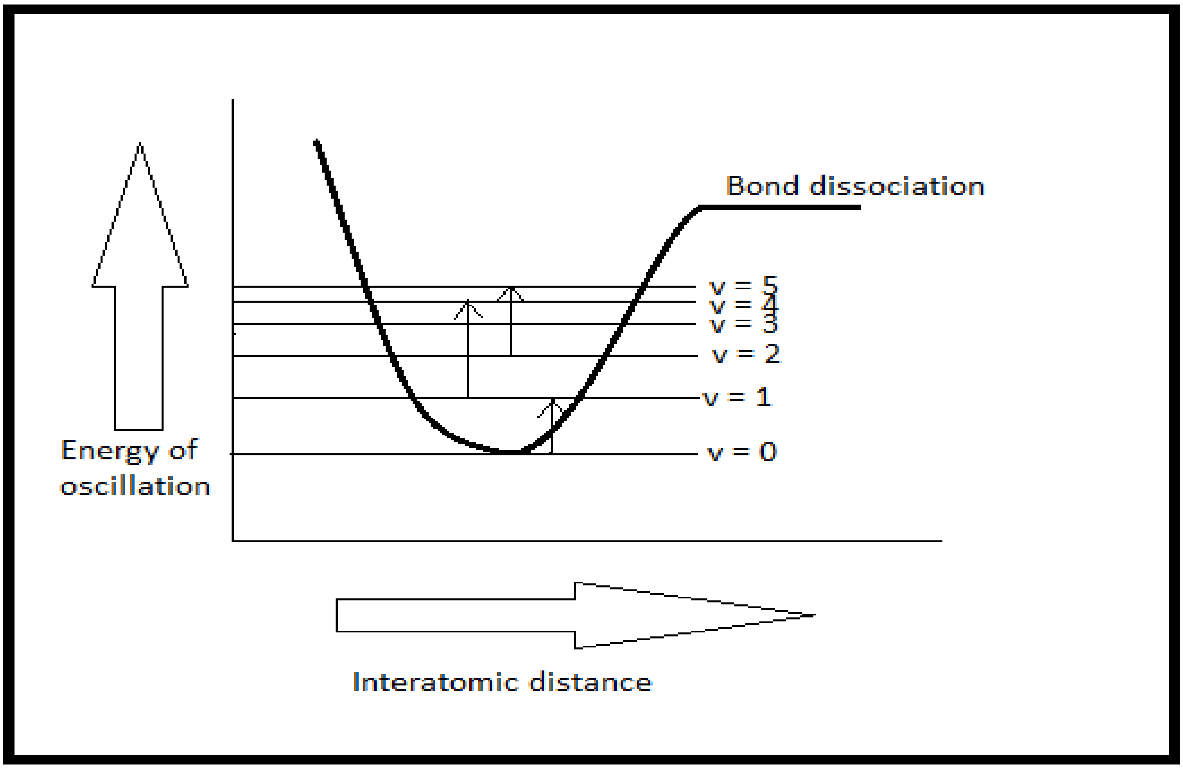


Figure 2.5: Anharmonic oscillation

Here the potential energy equation changes a bit here energy (E_{vib}) is going to be

$$E_{vib} = h*v(v + ½) – x*(v + ½)^2*h*v$$

Here the second term is added and x is called anharmonicity constant. The value of x ranges between 0.005 and 0.05. Now from the equation if the vibration energy level calculated then we can find that as the vibration level increases the energy gap between simultaneous energy gap between two adjacent level gets shorter from figure 2.5. In this model Δv is not restricted to +1 or -1 so molecule can jump from v = 0 to v = 2 or above depending on the energy it receives[2,5].

Taking Maxwell's distribution into consideration it can be stated that at ambient temperature most of the molecules are in vibration state v equals to 0 so when an electromagnetic radiation is passed through the sample most of the molecules starts to take the energy from the radiation and goes to higher vibration level. So the it can be clearly understood that the NIR spectra is dominated by the fundamental and overtone.

The intensity of an absorption band is dependent upon the magnitude of the dipole change during the displacement of atoms in a vibration. Both phenomena are present in great intensity associated with bonds involving the hydrogen atom and some other heavier element such as carbon, nitrogen and sulphur. The O-H, C-H, N-H and S-H bonds tend to present high anharmonicity and high bond energy with fundamental vibrational transitions in the region of 3000 – 4000 nm. Therefore, it allows to predict the overtones and combinations of the fundamental vibrations of such bonds to occur in the region of energy associated with NIR photons. Intensities are in between 10, for combinations, up to 1000, for successive overtones, times lower than the absorption resulting from fundamental vibrations[1,2].

2.6. NIR modes of Measurement

There are six different modes of measurement in NIR spectroscopy and those different sample presentation modes are an important factor in NIR measurements. The modes are diffuse reflectance, transmission, transflectance, interactance and transmittance through scattering medium. In the diagram below two of the most important modes are shown. Mostly in transmission mode transparent samples are measured in glass/quartz cuvettes, on the other hand solid samples are measured in Diffuse reflectance mode.

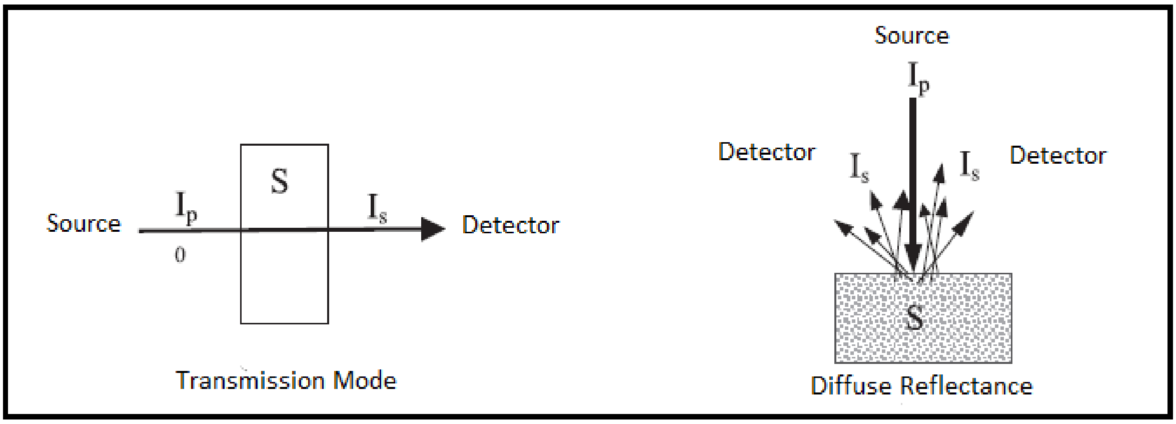


Figure 2.6: Transmission and diffuse reflectance mode

2.6.1. Diffuse Reflectance Mode

The diffuse reflectance technique is very unique because of its ability to interpret the spectra with minimal sample preparation as solid samples can be used directly in the mode. Scattering and absorbance by solid granules contribute to change the signal intensity. A rigorous treatment of the signal obtained in this type of measurement was established by Kubelka and Munk in 1931. This mathematical treatment results in the following equation that replaces the Beer´s law, while it has been earlier discussed that the Beer’s law cam be implemented only in transparent homogeneous materials where scattering is limited. Now the equation represents a relationship between concentration (C) and the diffuse reflectance (R).

$$\text{f(c)} = \frac{(1-R)^2}{2R}$$

In the equation R is the reflectance which can be given as,

$$\text{R} = \frac{I_R}{I_{R0}}$$

In the equation I_R represents the radiation reflected by the sample and I_{R0} is the radiation reflected by a non-absorbing material throughout the whole spectral range and their ratio gives us the reflectance. As it is very tough to collect and store the scattered radiation reflected along all the wavelengths this technique is not employed most often. Instead a more practical nonlinear equation is used for calculation of absorbance[1].

$$\text{f(c)} = \log\frac{1}{R}$$

Though the result obtained from the equation defers a little from Kubelka and Munk prediction but for a small change in reflectance assumed to present linear with the concentration of the sample analyte.

2.6.2. Transmission Mode

Mostly diluted sample solutions are used for this mode. Here the Beer's law is applicable and the equation for calculation of the transmission ratio is,

$$\mathrm{A} = \log\frac{1}{T} = \log\frac{I_s}{I_0} = abc$$

For a single wavelength,
I_0 Light intensity transmitted through an empty path
I_s is the light intensity transmitted through a sample(both I_0 and I_s are calculated for an equal distance path).
A is called the Beer-Lambert optical absorbance
a is absorption coefficient, cm
b is Path length (or sample thickness), cm
c is Concentration of absorbing material
T is Transmittance ratio.

There are four other techniques or modes for NIR spectroscopy but it can be easily verified that all those techniques fundamentally derived some way or the other from Transmission or Diffuse reflectance modes.

2.7. NIR Experimental Setup

The NIR laboratory set up used in this thesis work is DWARF-Star NIR spectrometer (StellarNet Inc., USA). The DWARF-Star NIR spectrometer is used to collect required data at a temperature of 25±0.5°C. The Peltier cooler system in this spectrometer is very essential feature for maintaining the temperature. As in this experiment absorbance or diffuse reflectance mode is used for that reason the solid sample needed to be finely sieved sample. The samples (finely sieved) for NIR analysis are placed on a standard quartz cuvette of 1 mm thickness. Quartz have ability of transmit radiation wavelength from 190 nm to 2500 nm which is very useful for this experiment. Cuvette with the sample in it was put upside down first for maximizing distance covered by the NIR radiation and it was also put in a way that covers the entire cross sectional area of the slit. This also results a smooth surface that assures a maximum reflection and thus a good signal-to-noise ratio. Four reflectance readings are taken from each specimen at three different angular positions. Also in between the repetitions the cuvette was properly stirred and it was noticed that at the surface of the cuvette the samples are properly spread. The quartz cuvette was wiped using a dry dust free tissue before each new sample reading. The absorption spectra consisted of wavelengths from 879 nm to 1755.55 nm and the interval was 1.75 nm. So in this range absorption calculated and recorded for 502 different wavelengths in abs format file.

For the experiment 5 replicate was taken for each angular position of the cuvette and for 4 different angular position a total of 20 data points taken and stored for each sample.

The NIRS needed calibration, and was first kept on for half an hour before calibration. Before taking any reading the instrument needs the bright spectra and the dark spectra for calibration. While keeping the source off a dark black tile was put on the slit and the dark spectra was recorded. Then putting the source on a white Teflon tile was used for recording the bright spectra.

The NIR data is collected using a bi-directional external optical fiber cable installed at three points, 120° to each other into the high intensity contact probe, an InGaAs detector array of 256 diodes and a tungsten halogen lamp to provide light source. The contact fiber is inserted at an angle of 45° to the plane on which the container is placed.

The schematic diagram for the NIR spectrometer used in this study is shown in figure 2.7.

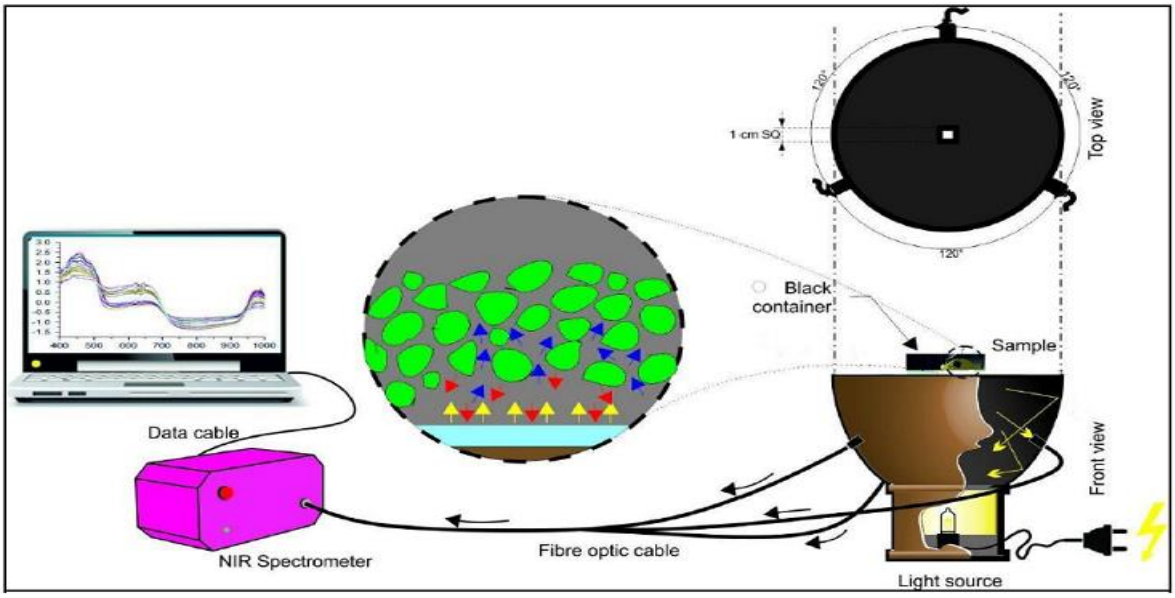


Figure 2.7: The schematic diagram for the NIR spectrometer

Source:

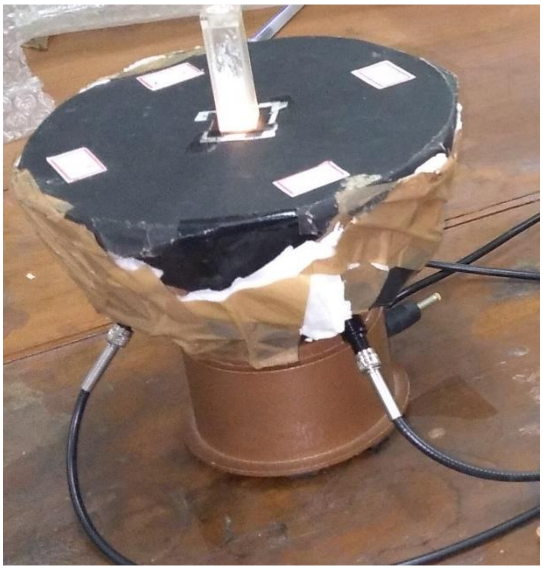


Figure 2.8: Tungsten Halogen light source

Specification:

- SL1 Tungsten Halogen light source has tungsten halogen lamp field with Krypton gas.
- 12V DC supply.
- Spectral range 350 to 2200 nm.
- Three optical fiber probe making 45° angle with the sample surface and making 120º angle among each other.

Detector:

Figure 2.9: Detector DWARF-Star

Specification:

- In GaAs detector array
- 900-1700nm wavelength range
- 512 pixel Resolving resolutions to 1.25nm.
- Integrated thermoelectric cooler.

References

1. C. Pasquini, “Near Infrared Spectroscopy: fundamentals, practical aspects and analytical applications,” *J. Braz. Chem. Soc.*, vol. 14, no. 2, pp. 198–219, Apr. 2003, doi: 10.1590/S0103-50532003000200006.

2. “A guide to near-infrared spectroscopic analysis of industrial manufacturing processes.”

3. Book “Principles of Instrumental Analysis’’ by Douglas A. Skoog (Stanford University), F. James Holler (University of Kentucky) and Stanley R. Crouch (Michigan State University).

4. *Tools of radio astronomy*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009.

5. B. Pecori, G. Torzo, and A. Sconza, “Harmonic and anharmonic oscillations investigated by using a microcomputer-based Atwood’s machine,” *Am. J. Phys.*, vol. 67, no. 3, pp. 228–235, Mar. 1999, doi: 10.1119/1.19230.

CHAPTER THREE Data management and analysis

3.1 Introduction

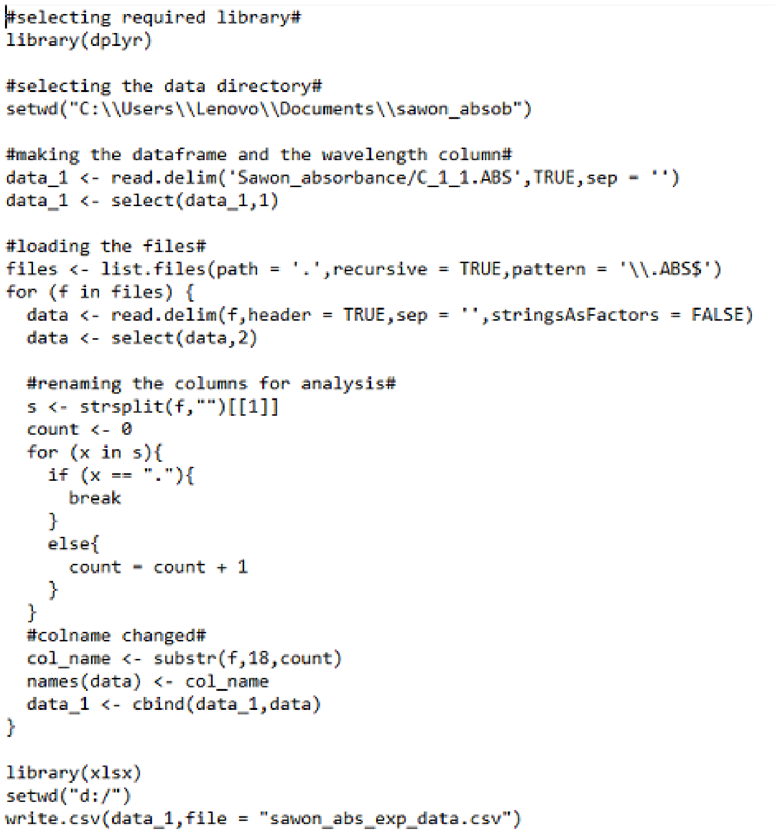
In this chapter All the methodologies used for the thesis starting from Data collection and storing till data analysis and classification algorithms are described. These techniques are often used in analysis of plant and agriculture products after NIR absorption mode data collection is done. All the methods were done either on R software or on MATLAB. Pre-processing is a very vital step of any data analysis then comes feature reduction techniques.

3.2. Sample Preparation and data collection

Sample preparation is very essential and important step in this thesis as NIR diffuse reflectance mode is used in this experiment. It was already discussed in the previous chapter that for agricultural products and medicinal products NIR reflectance mode is useful and in that mode there is minimal sample preparation efforts are required. It can be understood that the spectrometer response is better when near infrared radiation interacts with greater number of sample particles. So for this the sample needed fine powdered.

Six sample from six different places were collected. Then the samples were dried. After that the sample with stem and leaves cut into small pieces. A grinder machine is used for making dust from the sample. With butter paper and sieve fine sample powder was collected and ready to put into the cuvette.

Spectrawiz software was used for the data collection. 20 repetitive measurements were taken form one sample while the angular position of the cuvette was changed after 5 consecutive data collection. The data was in dot abs format basically it was in text format. So the stored data needed to be put in an excel file or in a csv file to use that for further pre-processing and calibration. In figure 3.1 the snap shot of the program used for data management is shown.



```
#selecting required library#
library(dplyr)

#selecting the data directory#
setwd("C:\\Users\\Lenovo\\Documents\\sawon_absob")

#making the dataframe and the wavelength column#
data_1 <- read.delim('Sawon_absorbance/C_1_1.ABS',TRUE,sep = '')
data_1 <- select(data_1,1)

#loading the files#
files <- list.files(path = '.',recursive = TRUE,pattern = '\\.ABS$')
for (f in files) {
  data <- read.delim(f,header = TRUE,sep = '',stringsAsFactors = FALSE)
  data <- select(data,2)

  #renaming the columns for analysis#
  s <- strsplit(f,"")[[1]]
  count <- 0
  for (x in s){
    if (x == "."){
      break
    }
    else{
      count = count + 1
    }
  }
  #colname changed#
  col_name <- substr(f,18,count)
  names(data) <- col_name
  data_1 <- cbind(data_1,data)
}

library(xlsx)
setwd("d:/")
write.csv(data_1,file = "sawon_abs_exp_data.csv")
```

Figure 3.1: snapshot of the program used for data handling

3.3. Chemo-metrics

Though the radiation beyond the visible radiation was present throughout the first decade of eighteen hundred but there was no significant advancement during the whole of eighteen hundred as the early literature suggests that transmission mode of measurement was used often in that ere. After 1930 the abortion mode was accepted as a standard for agricultural and plant specific measurement and quality assessment. That faces a new kind of situation as the NIR spectra was overlapping and it was difficult to interpret. Mainly the vibrations from –CH, -OH, -NH, -SH bonds are responsible for NIR spectra but all absorption bands are combination of fundamental and overtones that are very difficult to separately analyse. Now a day the evaluation of chemo-metrics proved very effective solution for those kind of problems[1].

Chemo-metrics is a multidisciplinary approach that can be used for qualitative and quantitative analysis for NIR spectra. Chemo-metrics involves statistical and mathematical approach to get relevant information from the spectra. That information then used for decision making[1]. The most common chemo-metrics methods implement the below mentioned steps:

- Measuring and collecting data
- Pre-processing data
- Multivariate analysis
- Calibration
- Validation

By practicing this few techniques quantitative and qualitative information from the spectral data is collected. In this experiment three main methods are used for analysis.

- Mathematical pre-treatment is used to enhance the information relevant for the study and reduce the influence of noise and other disturbances to increase the signal to noise ratio. Classical pre-processing methods are smoothing and normalization or standardization.
- Classification models are used to group similar category of information together. For that in this experiment a separability index is calculated and for different pre-processing their separability was checked to quantify the pre-processing techniques. Then classification models were implemented and it was checked that the models can differentiate the different classes from the NIR pre-processed spectral data. In that way a further study is done and the patterns that are formed those are validated by using regression method.
- The pre-processed data is quantified by using regression method.

3.4. Pre-processing

In this thesis there are several pre-processing techniques have been used. They can be classified into three groups one is smoothing, the second one is normalization and the third one is scatter correction technique. For smoothing in this experiment detrending technique has been used and for normalization standard normal variate (SNV) and max min standardization techniques have been used. On the other hand, multiplicative scatter correction technique is done upon the raw data[1,2].

3.4.1. Detrending

In a raw data there may be a trend that the data follows all throughout the wavelengths. So that trend may hamper the information that is relevant. Raw data may follow a liner trend or it may follow a nonlinear trend. A linear trend may indicate typically indicated a systematic increase of decrease. This is very much relatable to sensor drift. In this method to find the linear tend, the best fit line is calculated in least square sense for each of the sample data vector. Then for every wavelength the trend value is subtracted from the real raw data value to get the detrended values[3,4]. For detrending, separate data columns are separately treated. Sometimes nonlinear trends are present in a data at that time it is useful to try and fit a polynomial function to the data for that a quadratic equation may be considered and using the root mean square error minimization a proper fit is found.

So for this thesis for 120 sample data columns 120 separate trends are calculated. A linear trend has been found for each data vector separately using least square technique and the trend values were subtracted from the data vectors.

The linear trend is the best approximation of the data point in least square sense so if we need to approximate the equation:

$$y \sim Ax$$

y here is the data points

x is the input

A is the coefficient matrix.

Now if we put the least square equation that is going to be:

$$\min \|y - Ax\|^2$$

if above mentioned equation or objective function can be minimizes using Karush Kuhn tucker condition and this will yield the best fit line (Y) for the particular data vector. then after detrending the data will be:

$$Y_{det} = y - Y$$

Here Y_{det} is the detrended data vector.

3.4.2. Standardization

In this thesis there are three standardization techniques have been used. SNV and max min standardization is very popular while working with the NIR spectra with that for this thesis for data pre-processing the mean centering technique has also been used.

3.4.2.1. Mean Centering

There are many factors that hamper data analysis among them one of the most important is multicollinearity. Multicollinearity comes when there are multiple vectors present in a data which are linearly dependent. If X1, X2,…Xn are data vector in a particular data set. If

K1 * X1 + K2 * X2 +….+ Kn * Xn = 0, for K1 != 0, K2 != 0,…Kn != 0

K1,K2,…Kn are constant and non-zero

Then we can say that there is multicollinearity present in the data. So it can be understood when independent variables are highly correlated with one or more no of variables that then undermines the statistical significance of one or more independent variables. Multicollinearity can lead to bias interdependency among parameter estimates as well there can be inflated standard error. It is more problematic while finding the regression coefficients. Mean centering is done by calculating mean for each and every data columns and then subtracting that particular mean from its respective data column.

$$\mu_j = \frac{\sum_i X_{ij}}{n}$$, Xij is the data value at ith row jth column

N is the length of jth column. And mean of the jth column is μ_j. Now as the mean is calculated now the mean centered data is going to be.

$$X\_meancenter_{ij} = X_{ij} - \mu_j$$, for all i.

$X\_meancenter_{ij}$ is the mean cantered data.

3.4.2.2. Z-Score Standardization

In case of spectroscopy there is scattering present in each of the samples. These scattered radiations maybe different for different spectra as well there may be a multiplicative effect present in a spectroscopic output for scattering. This scattering effects may prevent required information collection from the spectral dataset. For reducing this effects, a well-known normalization is used that is sometimes called Z-Score Standardization or SNV (Standard Normal Variate) normalization[5]. Some times for getting a PCA, SNV is done on the data to scale in a particular unit. This is a very useful technique for comparison between two different data vector.

Mathematically it can be done by calculating the mean and standard deviation of every data separately then subtracting the mean from the respective data vector after that data vector is divided by its respective standard deviation.

$$sd_i = \sqrt[2]{\frac{\sum_j (X_{ij} - \mu_i)^2}{n-1}} \quad \text{,here } \mu_i = \frac{\sum_j X_{ij}}{n} \text{ is the mean of ith data}$$

Now for X_snv_{ij} ,

$$X\_snv_{ij} = \frac{X_{ij} - \mu_i}{sd_i} \quad \text{,for all j and i.}$$

X_snv_{ij} is the data matrix after applying SNV normalization.

3.4.2.3. Max_Min Standardization

In case of NIR Spectra the scattering is different for every repetitions and in case of absorption mode usually the ratio of two intensities are considered and the calibrated data is taken as absorption output but due to the difference in scattered radiation there will be difference in output scale. So for that reason scaling will take a very important part while analysing the data. In case of max min standardization, the data values are scaled between 0 and 1. That is the reason that comparison among the data vectors becomes easy.

For this thesis each data vector was taken and its maximum and minimum value was calculated and then the maximum value was subtracted from each value of the respective data vector and then it was divided by the value maximum minus minimum to get the max min standardization.

$$X\_maxmin_{ij} = \frac{X_{ij} - \max(X_j)}{(\max(X_j) - \min(X_j))} \quad \text{,for all j and i.}$$

Here Xij is the data at row i and column j. X_maxmin gives the data matrix after max_min standardization.

3.4.3. Multiplicative Scatter Correction

The basis of data analysis in NIR Spectroscopy is the scattering of light through the sample. As we previously discussed that scattering varies for different sample spectra. This difference also dominates the interpretations that can be drawn from the data vectors. In most of the applications it creates an absorbance shift in the spectrum that makes the chemical interpretation difficult. Also when measured in highly scattering medium difference in optical path length also creates an issue. Now multiplicative scatter correction reduces the mentioned problems as well it can eliminate the error due to multiplicative and additive scatter effects. It is also a very efficient technique that reduces the error due to slope changes in the sample data. MSC also improves the linearity in the NIR Spectroscopic data[6].

For doing MSC first the mean value of absorbance was calculated for each of the wavelengths, now taking the mean vector into consideration a regression line is predicted taking the raw data vectors as output.

$$X_i = a_i + b_i * X_{mean_i} + \text{error.}$$

Here the X_{mean_i} is the mean calculated and taken as reference input for the best fit line. a_i, b_i are the regression coefficient. For this calculation in this thesis the least square solution is used. After getting a_i, b_i to calculate the MSC data first a_i is subtracted from the respective raw data vector and then the result is divided by b_i.

$$X\_msc_{ij} = \frac{X_{ij} - a_i}{b_i} \quad \text{, for all i and j.}$$

3.4.4. Principal Component Analysis

The near infrared spectrometer provides absorbance values for wavelengths 900 to 1700 Nano meter with 1.75 Nano meter step changes. Those can be treated as the features for the data set. Taking this into account there are more than five hundred features that are responsible for the spectral data analysis. As there are a huge no feature present in the data if any machine learning algorithm is trained on this data that inevitably lead to under fitting. While the feature is reduced the required information may also be lost. This is why for this thesis Principal Component Analysis plays a very important role. PCA reduces the dimensionality while preserving the variability and statistical information[1,7]. PCA is an orthogonal transformation technique. Orthogonal transformation is a linear transformation that preserves the inner product of the basis vectors. The inner product tells about the angels between the vectors and their length. For example if u and v are two vectors and their inner product is <u , v> then if those vector are transformed orthogonally and if T:V -> V is the transformation. Then <u , v> must be equal to <Tu , Tv>. Tu and Tv are the transformed vectors. Taking X as the data matrix PCA tries to orthogonally transform the data and gives the direction in which variance is maximum as the first principal component and the second maximum variance direction as the second principal component and so on. So if V1 is the direction in which the variance is maximum then mathematically

$$\text{Variance} = \frac{1}{(n-1)} * \sum_{i=1}^{n} (V1^T * X_i)^2$$

As the ith component's projection on V1 is given by $V1^T * X_i$.

The equation is to be maximized to get the first principal component subjected to a constraint $||V1||$ = 1 due to the fact that PCA is needs to prevent the inner product of the basis vectors. The above equation can be reduced to $V1^T R\, V1$ after simplification. Here R is the covariance matrix for X. now using KKT condition for maximization the solution is going to be

$$R\ V1 = \lambda\ V1 \quad , \lambda \text{ is the Eigen value of R.}$$

And putting this in to the equation max ($V1^TR\,V1$) yield.

$$V1^TR\,V1 = V1^T\lambda\,V1$$

As λ is a scalar quantity the equation can be written as:

$$V1^TR\,V1 \ = \lambda\,V1^T\ V1 = \lambda\ ||V1||.$$

From the solution it can be stated that if λ is maximized the whole equation and the variance of the projected vectors in the V1 direction is also maximized.

After getting the required directions if the data vectors are projected upon them that gives the required principal components.

Formally the covariance matrix is calculated at first. Then the eigenvalues of the matrix (R) is found then the eigenvalues are sorted in descending order and then their respective eigenvectors are also found out and put in order. Then multiplying that eigenvectors and the data matrix yield the principal components.

$$R_{ij} = \ \frac{1}{(n-1)}\frac{1}{(m-1)}\sum_{i=1}^{n}\sum_{j=1}^{m}(X_i-\mu_i)(X_j-\mu_j)^T$$

R_{ij} is the covariance matrix and if its eigenvector matrix is found to be V then the PCA matrix is V^TX .

Now checking the variance that the principal components can explain from the λ (eigenvalues) required number of principal components are chosen for further analysis.

3.5. Separability Index

This thesis is based on separability of Swertia chirayita of different geographical location after PCA in this thesis there is a need for quantification of separability of different classes. For this purpose, separability index calculation is most common and frequently used technique[8,9]. More the separability less complex classification model that can be trained using the data. Actually separability estimates the average number of instances that have a nearest neighbour with same label. In this after feature reduction separability index was calculated and according to that quantification a decision can be taken that which of the classification algorithm can be used for this particular case. As there are eight different pre-processing technique was used in this project for each of those pre-processing separability index was calculated and quantified.

The method of calculating separability index involves two scatter matrices one is within class scatter matrix and another is between class scatter matrix. After calculating both of them if the trace of those matrix is divided that would give the separability index[9].

The between class separability matrix(S_b) can be calculated using the below mentioned formula.

$$S_b = \sum_{i=1}^{c} n_i(m_i - m)(m_i - m)^T$$

Here c is the number of classes for this thesis the number of classes are six and n_i denotes the number of classes in ith class. m_i is the mean vector of samples in the ith class and m is the overall mean vector taking all the classes together.

Now the within class separability matrix S_w is calculated by,

$$S_w = \sum_{i=1}^{c} \sum_{j=1}^{n_i} (X_{ij} - m_i)\left(X_{ij} - m_i\right)^T$$

Here Xij is the ith row and jth element of the data matrix and j is iterated within ith class from 1st data to the last data $n_i$in the ith class.

After calculating these two matrix their trace is calculated and divided from each other to get the separability index(J).

$$J = \text{trace}(S_b)/\text{trace}(S_w)$$

3.6. Prediction using PCR

PCR is a regression method were despite using the feature (wavelengths for NIR absorption mode) the principal components after doing PCA is used. The number of principal components used to train the regression or prediction equation may vary depending upon the ability of the principal components to express the variation in the data. Often 95 percent variance is taken as a standard for selecting principal components. In this study four principal components were taken for PCR[1,2].

For NIR spectral data analyses one of the main problems is calibration. Constructing an equation for the relationship between the measured spectral values and the chemical composition. Now in ordinary least square technique there may be multicollinearity present in the features. It can be seen that correlation for two or more data features may reach 80 to 90 percent then training a regression model becomes difficult.
The prediction equation will be:

$$y = b_0 + X^T b + \text{error}$$

X is the data matrix. b_0 and b is the intercept and coefficients respectively. For getting the prediction equation the least square solution is used. If y is the output, then the predicted output is going to be.

$$y_{pred} = (X^T X)^{-1} X^T y$$

Taking the output y and the data X , y_{pred} is found.

The prediction error or residual standard error is calculated by using the predicted output and the actual output. The error is called mean square error.

$$MSE = \sqrt[2]{\sum\nolimits_{all\,y} (y - y_{pred})^2}$$

In this study four principal components are used to formulate the linear regression. Then from the predicted output the error was calculated.

References

1. Y. Roggo, P. Chalus, L. Maurer, C. Lema-Martinez, A. Edmond, and N. Jent, “A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies.,” *J. Pharm. Biomed. Anal.*, vol. 44, no. 3, pp. 683–700, Jul. 2007, doi: 10.1016/j.jpba.2007.03.023.

2. K. Héberger, “Chemoinformatics—multivariate mathematical–statistical methods for data evaluation,” in *Medical applications of mass spectrometry*, Elsevier, 2008, pp. 141–169.

3. E. Benes, M. Fodor, S. Kovács, and A. Gere, “Application of detrended fluctuation analysis and yield stability index to evaluate near infrared spectra of green and roasted coffee samples,” *Processes*, vol. 8, no. 8, p. 913, Aug. 2020, doi: 10.3390/pr8080913.

4. K. E. Jang, S. Tak, J. Jung, J. Jang, Y. Jeong, and J. C. Ye, “Wavelet minimum description length detrending for near-infrared spectroscopy.,” *J. Biomed. Opt.*, vol. 14, no. 3, p. 034004, Jun. 2009, doi: 10.1117/1.3127204.

5. L. Xu, P.-T. Shi, Z.-H. Ye, S.-M. Yan, and X.-P. Yu, “Rapid analysis of adulterations in Chinese lotus root powder (LRP) by near-infrared (NIR) spectroscopy coupled with chemometric class modeling techniques.,” *Food Chem.*, vol. 141, no. 3, pp. 2434–2439, Dec. 2013, doi: 10.1016/j.foodchem.2013.05.104.

6. Isaksson T, Næs T. The Effect of Multiplicative Scatter Correction (MSC) and Linearity Improvement in NIR Spectroscopy. *Applied Spectroscopy*. 1988;42(7):1273-1284. doi:10.1366/0003702884429869.

7. I. T. Jolliffe and J. Cadima, “Principal component analysis: a review and recent developments.,” *Philos. Transact. A Math. Phys. Eng. Sci.*, vol. 374, no. 2065, p. 20150202, Apr. 2016, doi: 10.1098/rsta.2015.0202.

8. M. Yektaii and P. Bhattacharya, “A criterion for measuring the separability of clusters and its applications to principal component analysis,” *Signal Image Video Process.*, vol. 5, no. 1, pp. 93–104, Mar. 2011, doi: 10.1007/s11760-009-0145-0.

9. Molder, Cristian. (2004). Feature Extraction for Classification: A Survey I. Linear Methods. MTA Review. 71-76.

CHAPTER FOUR Detailed experimental discussion

Six sample collected from different geographical locations (Sikkim, Nepal) were dried and made powder to put into the cuvette for NIR analysis. in the figure 4.1 the whole system is shown.

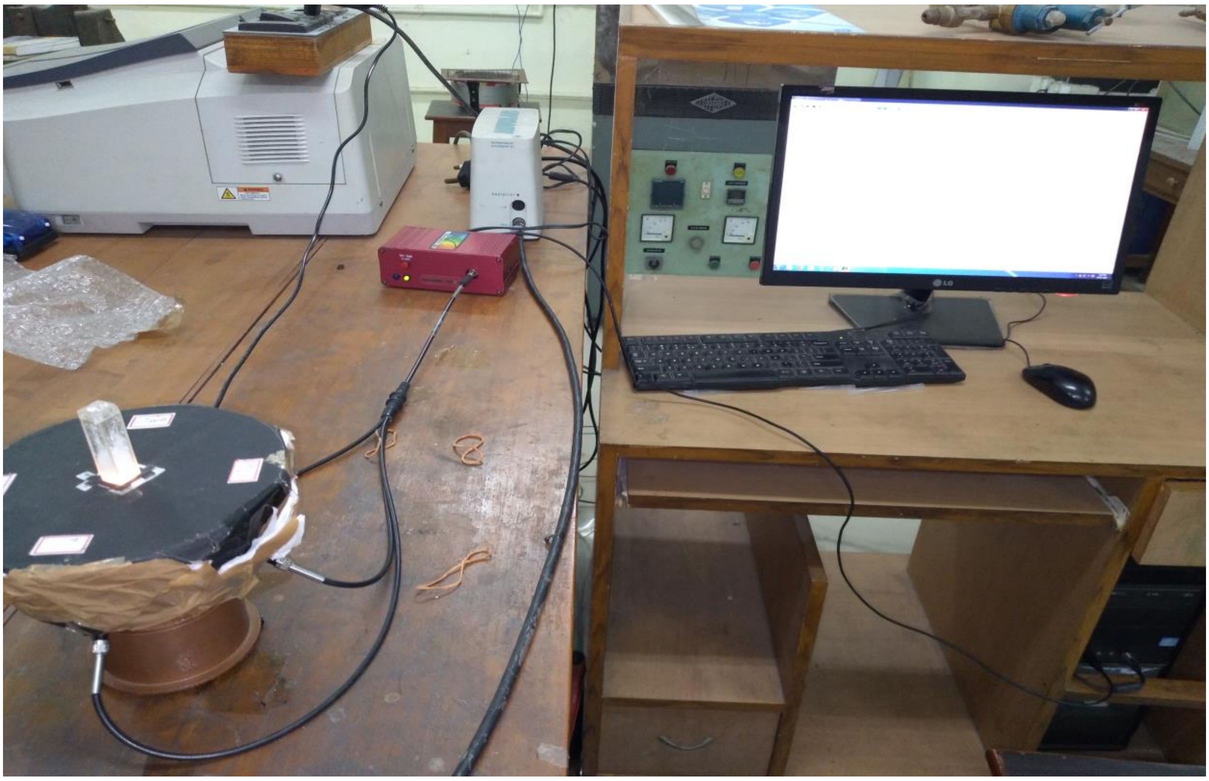


Figure 4.1: The whole sensor arrangement used for this study

The connections were made using USB cable. The source can be seen connected with the detector/spectrometer using fiber optic cable. After data collection that data is then converted into Excel format for analysis. and R and MATLAB software were used for classical analysis of NIR data. Taking the separability index values into consideration optimization was done on the wavelength ranges and it was found that the data within wavelength 1120nm and 1650nm is the most effective for analysis.

4.1. Raw data plot

Using MATLAB a line plot of raw data was done to analyse the peaks and trends present in the data. In the X-axis the wavelengths were plotted and in the y-axis the absorption values were plotted given if figure 4.2. The peaks at 1280nm and 1440nm can be seen for all the data.

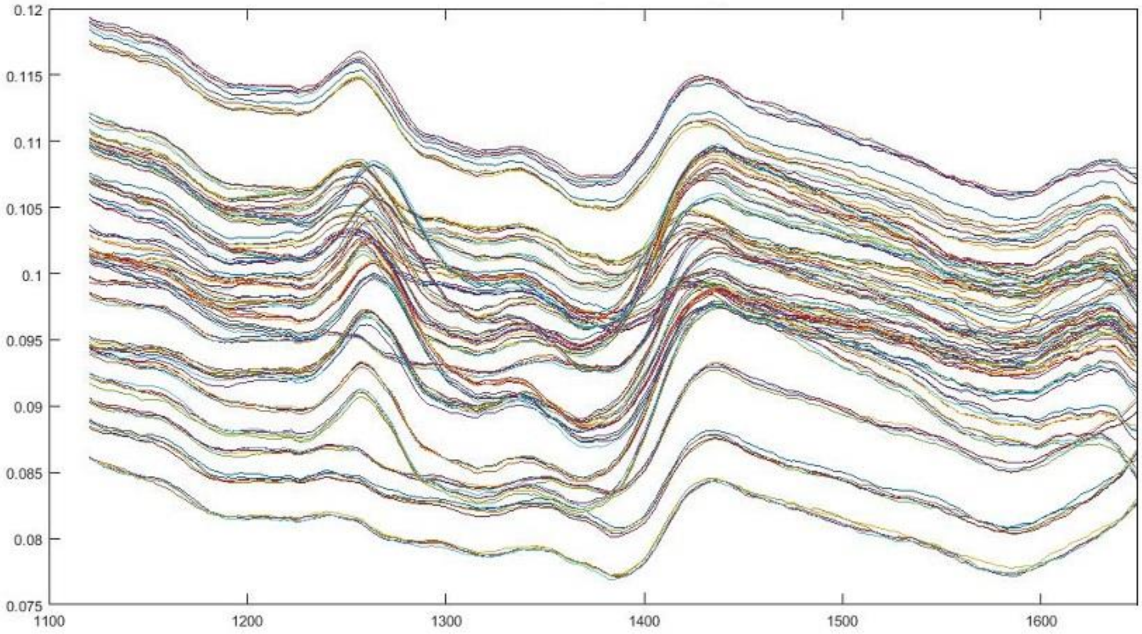


Figure 4.2: Line plot of raw data (X-axis wavelength, Y-axis absorbance)

Then principal component analysis was done and then PC1 and PC2 was taken for the scatterplot. It can be seen in the plot that PC1 can explain 91.8 percent of variance and PC2 can explain 6.9 percent variance can be seen in figure 4.3.

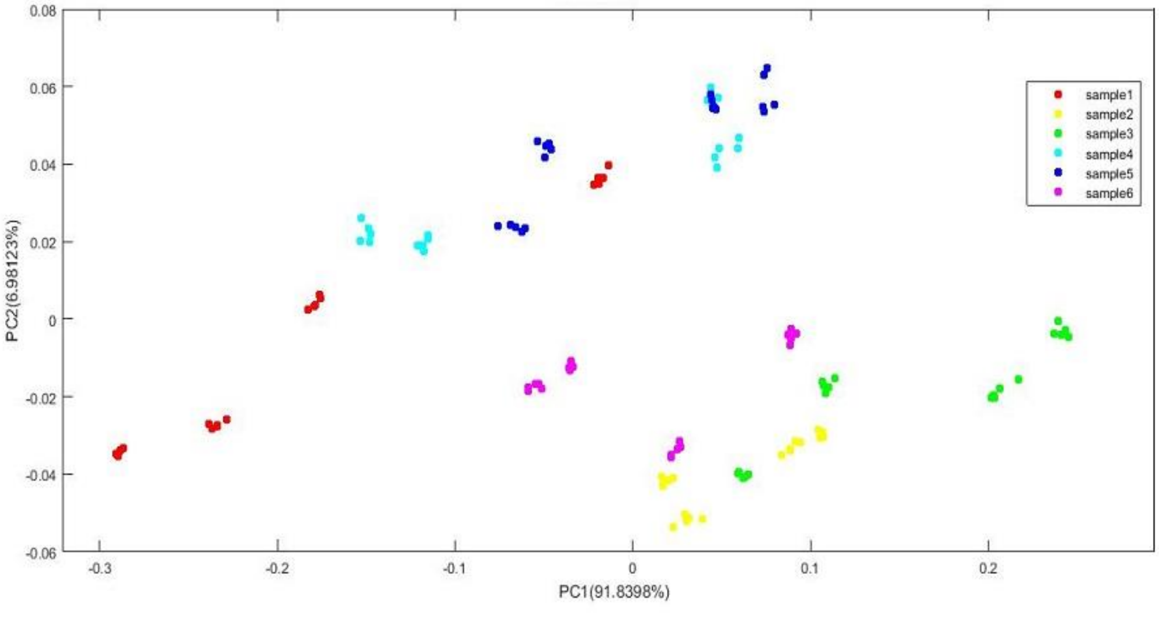


Figure 4.3: Raw data scatter plot (X-axis PC1, Y-axis PC2)

Also the samples form sample 1 to sample 6 is overlapping so separability is very less. For that reason, various pre-processing was don on the data.

4.2. Pre-processing

Seven different pre-processing was done on the raw data and their respective separability index values were calculated and then comments were made on that about the effectiveness of the pre-processing techniques.

4.2.1. Scatter correction

For implementing scatter correction in this study multiplicative scatter correction technique was used. In the previous chapter the detailed discussion about multiplicative scatter correction was given. Multiple scatter correction is implemented to correct the data from the errors that are added due to scattering of light radiation. The additive and multiplicative both the scattering effects can be rectified after using this technique. Using the least square solution, the regression coefficients were found then using that by using subtraction and division the MSC data was found in this thesis. Line plot given in figure 4.4.

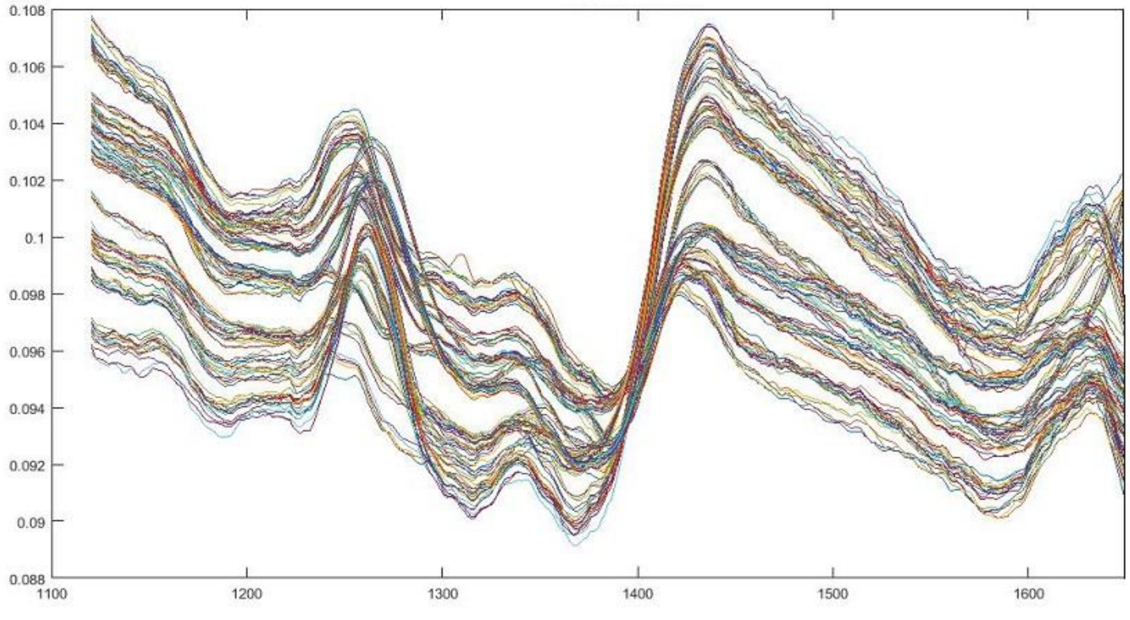


Figure 4.4: Line plot of MSC data (X-axis wavelength, Y-axis pre-processed-absorbance)

After scatter correction is implemented it can be seen that the data trends are somewhat subtracted from the raw data. Different samples are trying to occupy a certain trend but still the data had more than 300 wavelengths that are considered as features. So for further analysis one feature transformation technique was used(PCA) then taking two principal components a scatter plot was done to understand the sample separation.

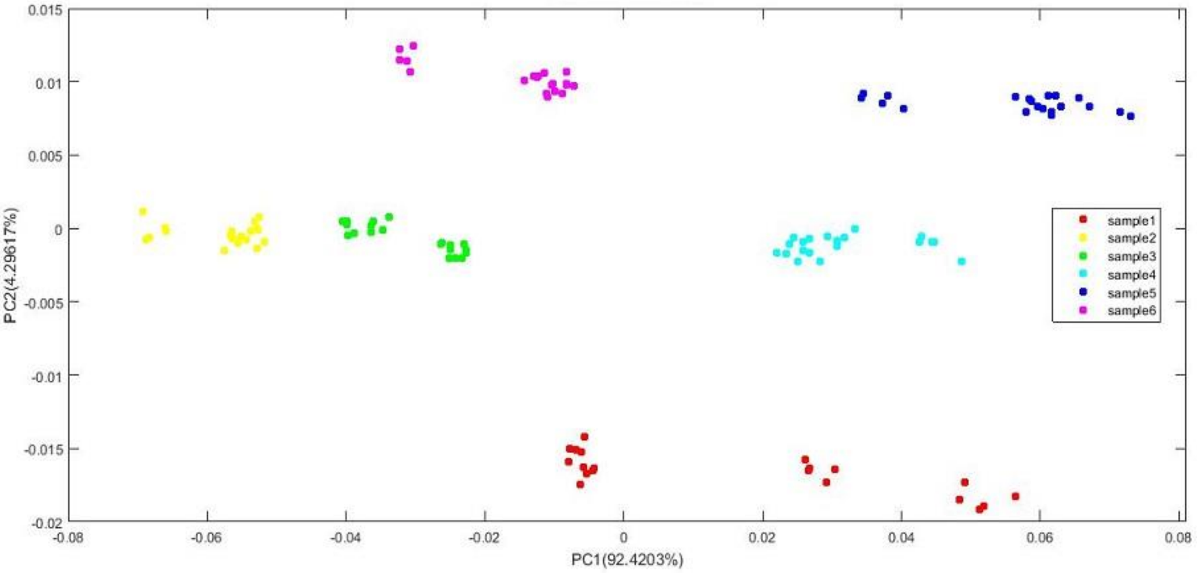


Figure 4.5: MSC data scatter plot (X-axis PC1, Y-axis PC2)

In figure 4.5, PC1 here expresses 92.4 percent of the variation in the data and PC2 can explain 4.2 percent variability. In this figure the sample clusters are more or less separable. Sample 1 is a bit scattered as sample 5. But sample 2 and 3 also sample 4 are following a certain behaviour and can be clustered easily. Relatively the separability value was also increased.

4.2.2. Z-Score standardization

This is widely used normalization technique. The discussion about this technique is given in the previous chapter. This technique also reduces the

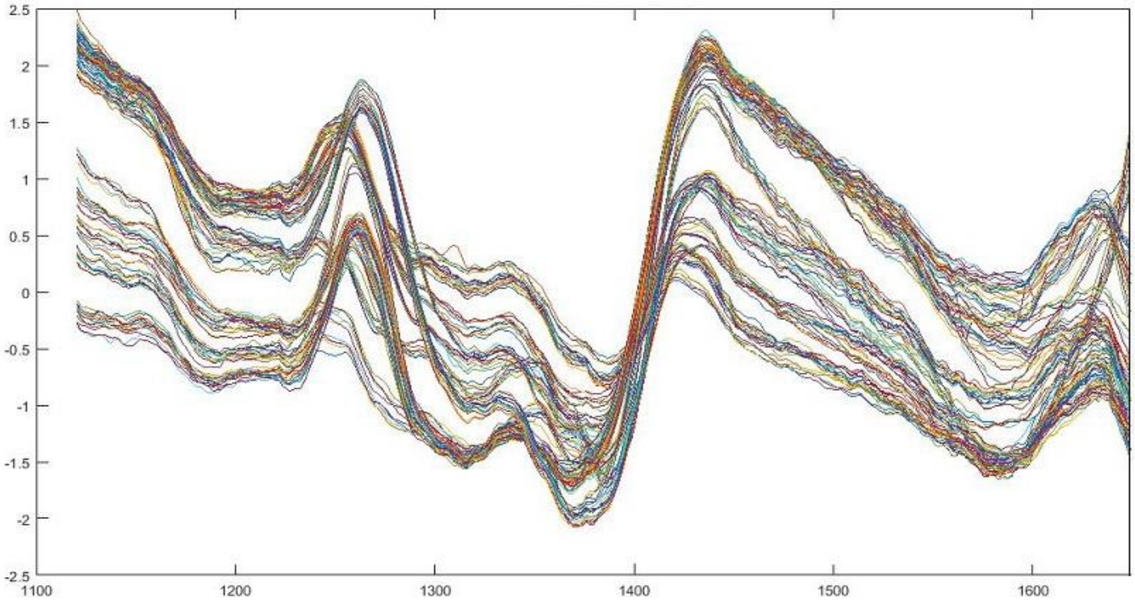


Figure 4.6: Line plot of SNV data (X-axis wavelength, Y-axis pre-processed-absorbance)

scattering effect for NIR spectroscopic data. Also it does a scaling on the data so that comparison between data samples are easier. Here also in the figure 4.6, a line plot of data is shown and x-axis represents the wavelength and y-axis represent the normalized absorption values. The peaks are also clear in this pre-processing and to understand the separability a feature transformation was done and a scatter plot is shown in the figure 4.7.

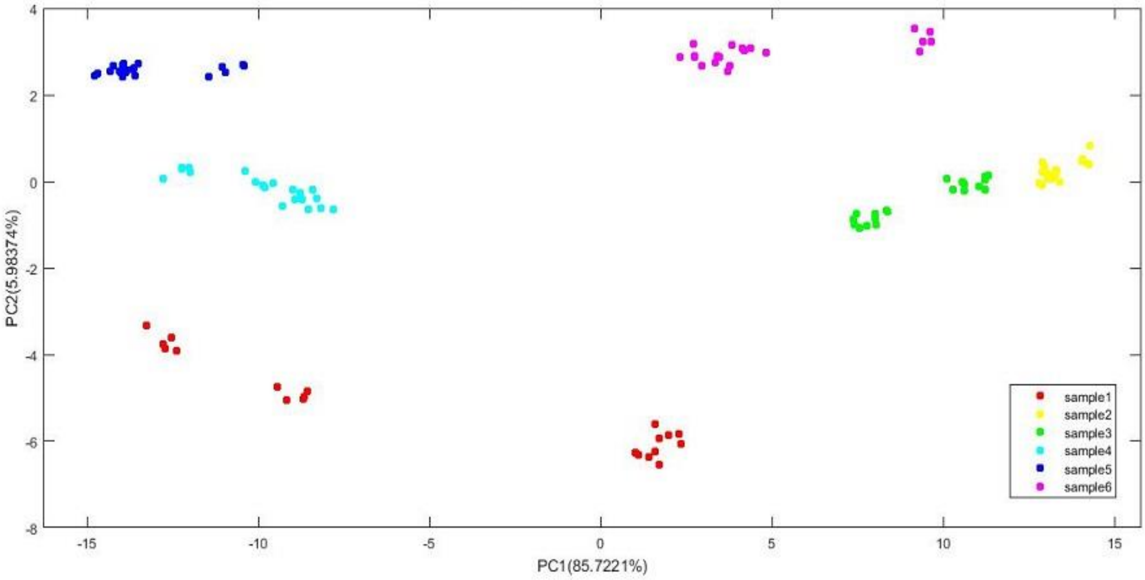


Figure 4.7: SNV data scatter plot (X-axis PC1, Y-axis PC2)

PC1 representing 85.7 percent variance and PC2 representing 5.98 percent variance sample 1 is scattered and other than that all the other samples following a particular trend and occupying a certain space in the plot. Separability is also improved from the raw data.

4.2.3. Mean centering

This technique is used to reduce the multicollinearity in the data features. That is important to analyse the data as the correlation among the data column is reduced. This technique is also discussed in the previous chapter. After this normalization all the data vector's mean becomes zero. This technique is also very widely used in the NIR spectral data pre-processing. Here in the figure a line plot of all the data vectors are plotted and it was compared with the other pre-processing data's line plots. X-axis represents wavelength and y- axis represents the mean-centered wavelength values in the figure 4.8.

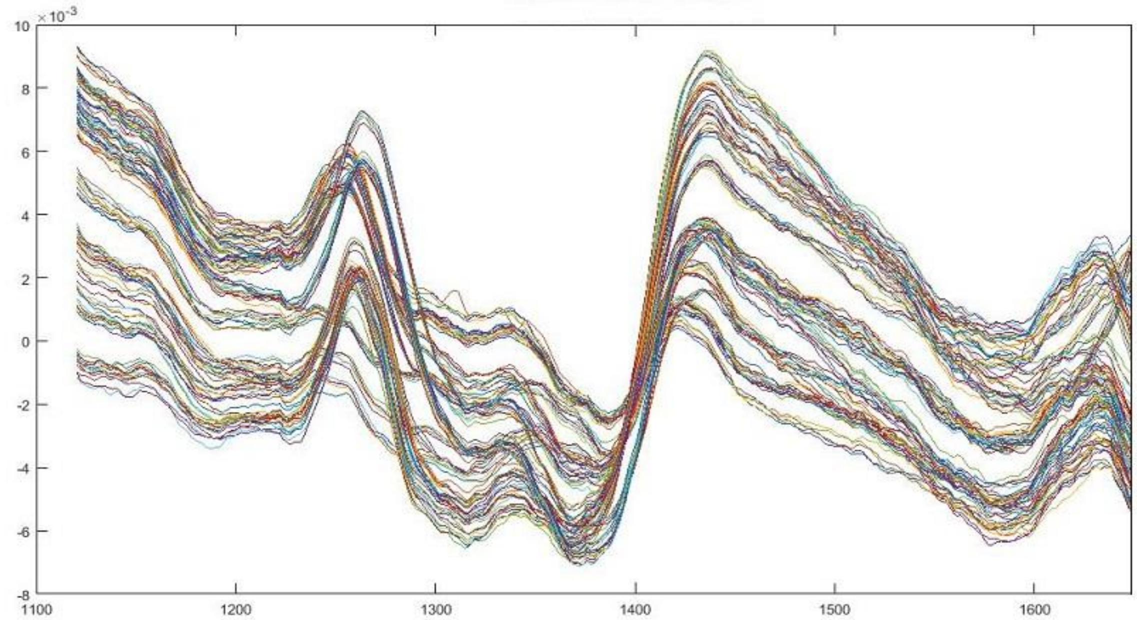


Figure 4.8: Line plot of mean centered data(X-axis wavelength, Y-axis pre-processed-absorbance)

Then the principal component analysis was done on the data and scatter plot given in figure 4.9, where PC1(var = 85.96) and PC2(var = 8.052) was plotted.

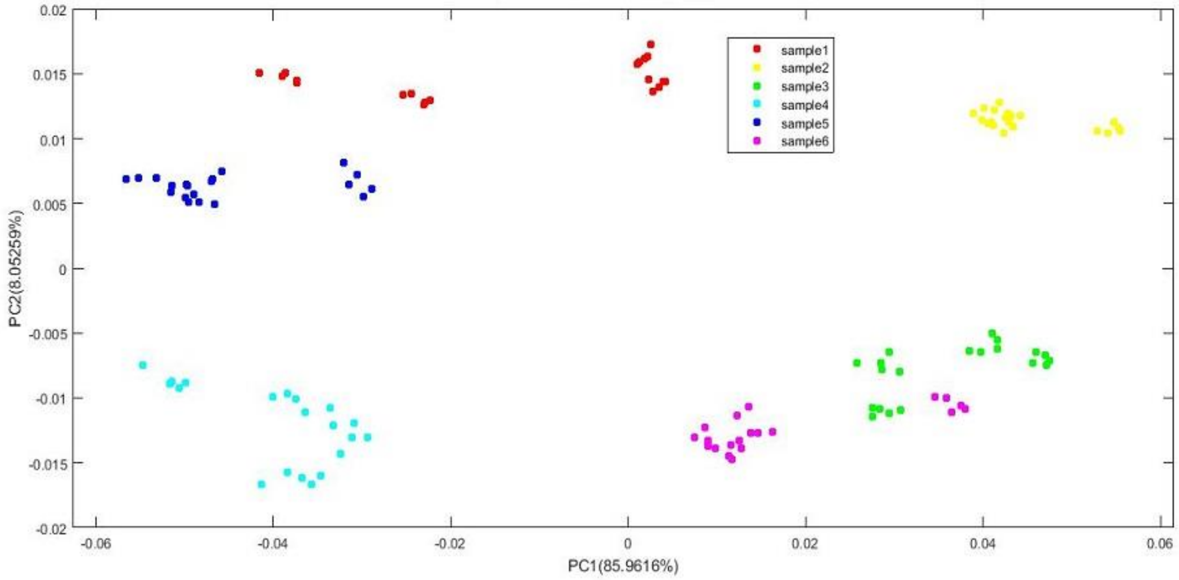


Figure 4.9: Mean centered data scatter plot (X-axis PC1, Y-axis PC2)

In this case all the samples are not separable sample 6 and sample 3 are overlapping. Though the other samples are fairly separable the separability index value found to be not high. Also sample 1 is more spread out in the plot.

4.2.4. Minmax standardization

Here also all the data were treated same way and the minmax standardized data was plotted with all the wavelengths in a line plot can be seen in figure 4.10, the PCA was applied and the principal components were plotted in a scatter plot.

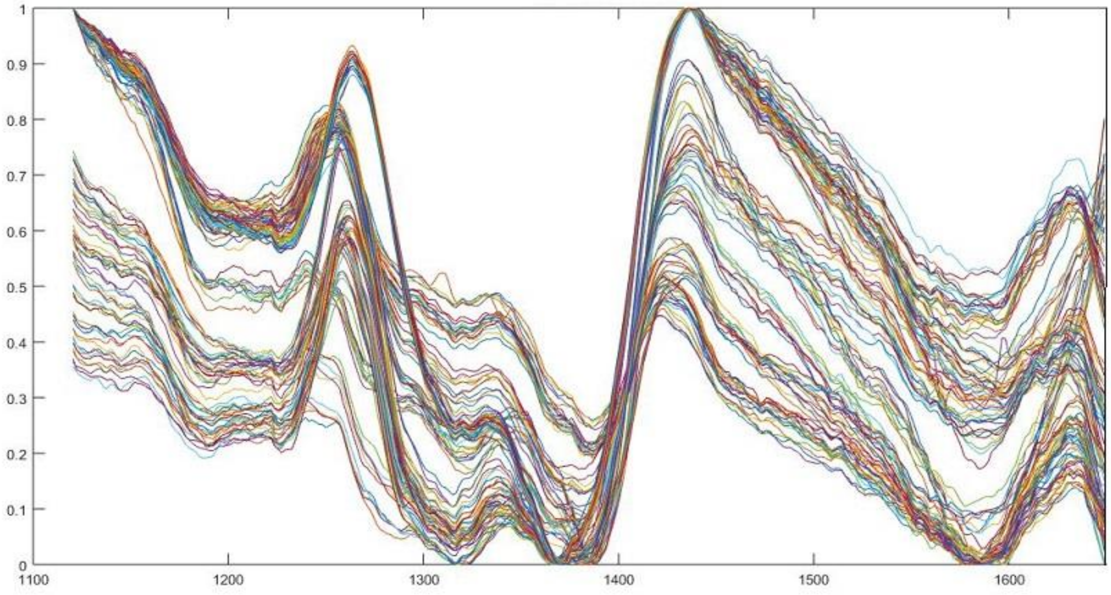


Figure 4.10:Line plot of maxmin standard data(X-axis wavelength, Y-axis pre-processed-absorbance)

This standardization scales the data in the range [0, 1]. Here also the peaks that concerned in this study is also very clear. Different samples following different trends.

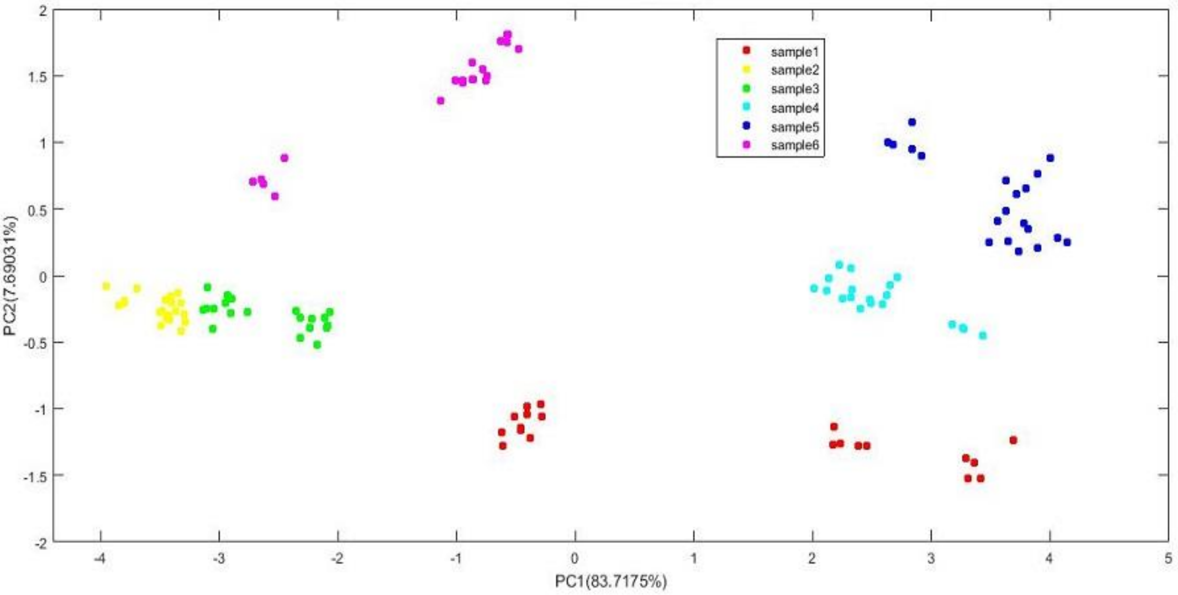


Figure 4.11: Maxmin standard data scatter plot (X-axis PC1, Y-axis PC2)

Both principal components together explain more than 90 percent of variability of the data in figure 4.11. Sample 2 and 3 are not clearly separable. Sample 1 is scattered other than that sample 4 and 5 are clearly separable.

4.2.5. Detrending

As discussed in the previous chapter for each of the sample data a linear trend was calculated and then that trend was subtracted from the actual data to get the detrended data.

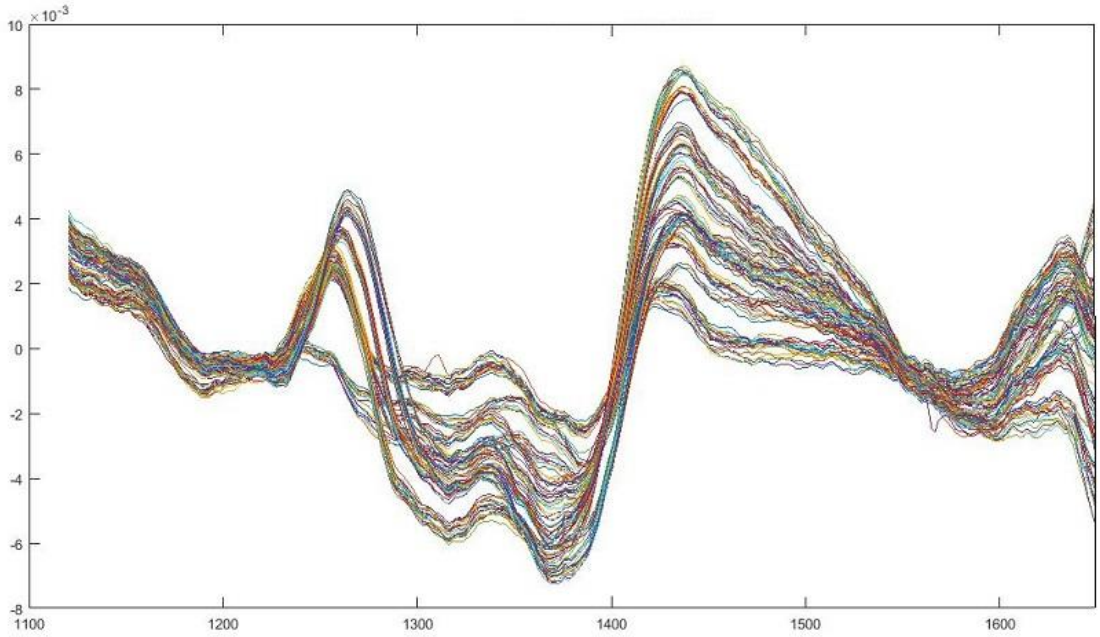


Figure 4.12: Line plot of detrended data(X-axis absorbance, Y-axis detrended absorbance)

it can be seen from the figure 4.12 that the data trends were removed and the peaks at particular wavelengths at 1250nm and 1339nm,1420nm are clearly visible and the absorbance values or the detrended absorbance vales at those peaks can be seen varying for different samples. Some of the samples are following a particular trend while one or two samples follow a different trend all together. For quantifying the detrended data principal component analysis was done and a separability index for that is also calculated and provided in the table below.

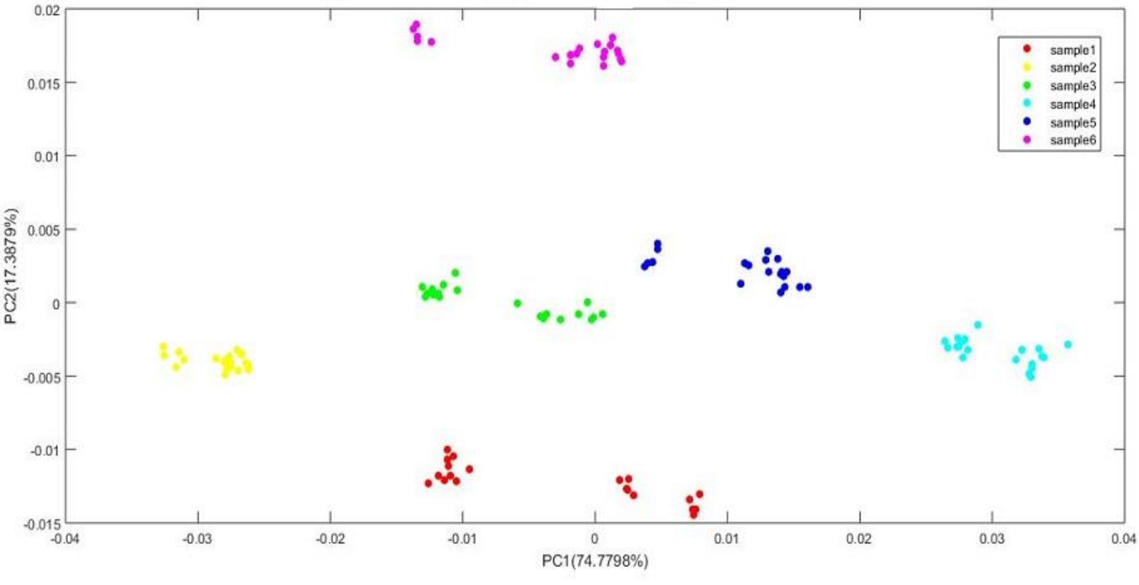


Figure 4.13: Detrended data scatter plot (X-axis PC1, Y-axis PC2)

The principal component PC1 can express 74.77 percent variation in the data and the second principal component PC2 can express 17.38 percent variation. Here the separation between the sample data is clearly visible so for that obvious reason the separability index value for this particular pre-processing. From this plot 4.13 it is seen that there is similarity in the phytochemical composition among samples. Data points of a certain sample is compactly placed.

4.2.6. MSC on Detrended data

This kind of compound techniques are also implemented often in NIRS data analysis. In this technique first the raw data was taken and detrending of the data was done. Then scatter correction technique was also implemented on that pre-processed data. Then the previously applied techniques were also done to analyse the data. There is a line plot shown in the figure 4.14 were x-axis represents the wavelength values and the y-axis represents the pre-processed absorbance values and in figure the Scatter plot of the two principal components were shown. As detrending gave the highest separability index value that’s why this particular compound technique was checked.

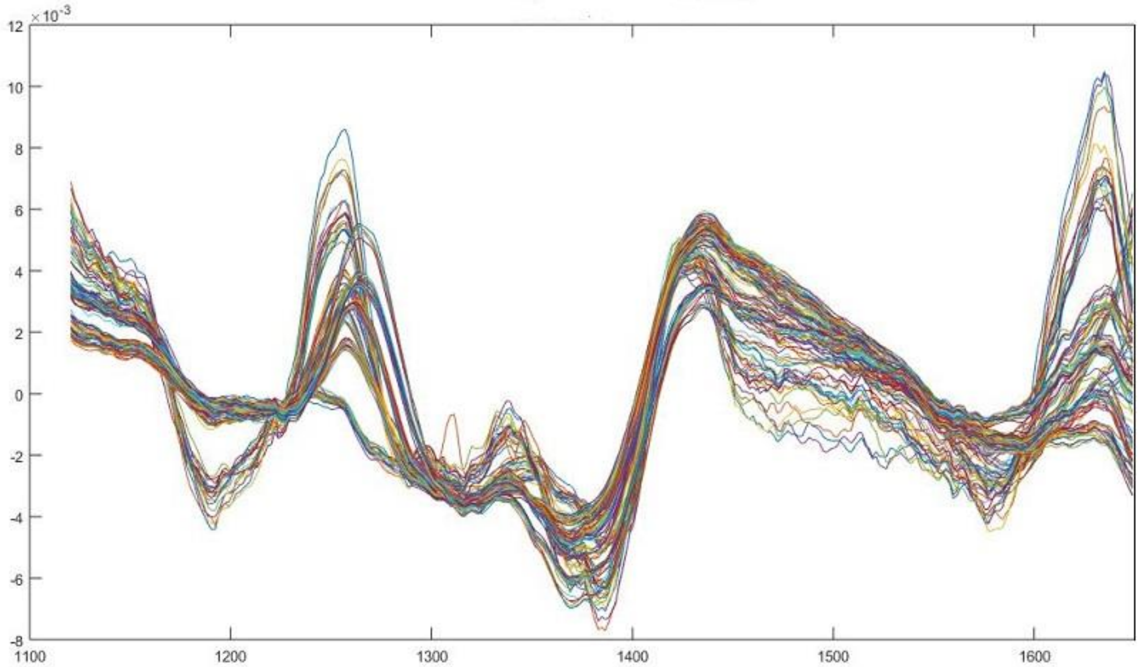


Figure 4.14: Line plot MSC on detrended data (X-axis absorbance, Y-axis scatter corrected absorbance)

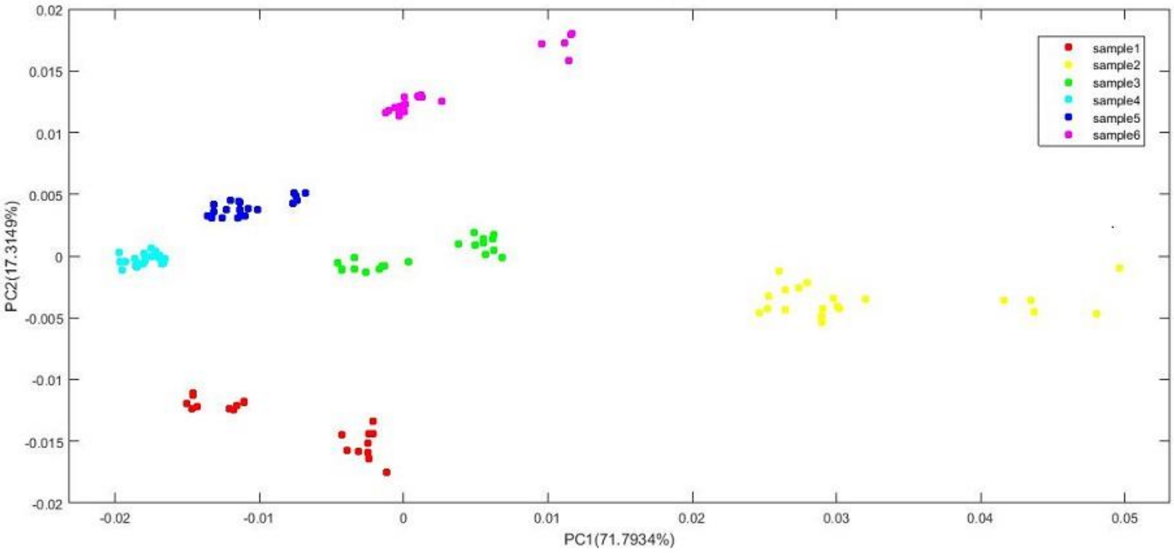


Figure 4.15: MSC on detrended data scatter plot (X-axis PC1, Y-axis PC2)

In the line plot figure 4.15 it was seen that the peaks are now shifted a small amount more or less each of samples following a certain trend and the variability at the peaks value was also observed. Here PC1 has a variability of71.79 percent and PC2 has variability of 17.31 percent. Also the samples are more or less separable. The separability values are given in the table.

4.3. Separability analysis

For discrimination separability index was calculated for each of the different pre-processed data. As discussed in chapter 3 two matrices within class scatter matrix and between class scatter matrix was calculated. Then using trace of that separability index was determined. For calculating the separability index four principal components were considered. In the table below the separability index values as well as PC1 and PC2 variability is given.

Table 4.1: Table of principal components variance explanation and separability for different pre-processed data

Pre-Processing techniques	PC1 variance explained	PC2 variance explained	Separability index
Raw data	91.84	6.9812	2.0217
Multiplicative scatter correction	92.42	4.2962	9.8198
Z-Score/ SNV normalization	85.722	5.9837	10.393
Mean Centered data	85.962	8.0526	12.411
Max min standardization	83.717	7.6903	9.6309
Detrended data	74.78	17.388	14.516
MSC on Detrended data	71.793	17.315	12.229

Form the table 4.1 taking the separability index values into consideration it can be stated that the most effective pre-processing technique is Detrending for this particular case. The mean centering technique also gave good accuracy while separability is concerned. Then comes standard normal variate normalization technique. MSC and min max standardization gives more or less similar separability.

4.4. Prediction analysis

Principal component regression is used for this study to understand the prediction accuracy of the data. For different pre-processing the prediction error was calculate and provided in the table. Also there are other parameters to understand the model validation those are also given in the table. In each of the pre-processing data four principal components were taken to train the regression model.

The number of total samples were divided randomly into two groups one for training and other for testing. 70 percent of the sample data (84) taken for training and rest 30 percent (36) used for testing and MSE error finding.

Table 4.2: Residual error and prediction error for PCR model for different pre-processing

Pre-Processing techniques	Residual Standard Error	R squared	Prediction error test data
Raw data	0.3615	0.9579	1.72024
Multiplicative scatter correction	0.1877	0.9886	1.231996
Z-Score/ SNV normalization	0.2938	0.9722	1.460319
Mean Centered data	0.1543	0.9923	0.7348995
Max min standardization	0.4793	0.9259	2.56541
Detrended data	0.13	0.9945	0.6279176
MSC on Detrended data	0.3053	0.9699	1.453096

Residual standard error was calculated using training data and the error after finding the best fit prediction line this is a standard cross validation error calculated on the training samples. The prediction error was found while prediction using the test data and all the errors are recorded in the table 4.2.

From table 4.1 and 4.2 it is evident that detrending is the pre-processing that is most effectively separates the data. Then comes the mean centered data. And the prediction as well as separability indicates that the bioactive chemical composition in those six samples are different. Through which the quantification and classification studies can be done on the data.

CHAPTER FIVE: **Conclusion and further scope**

Conclusion

Considering the impact of Swertia chirayita on human health and medicines as well as the economic implications of this herb in India and throughout the world there is an immediate need of research based quantification of this particular plant. This study takes up the opportunity. Simple and cost effective quantification through Near Infrared Spectroscopy was the aim of this study.

It is quite evident from this thesis that Swertia chirayita grown in different geographic location can be differentiated through the NIR spectral data. This study reveals that detrending the sample data collected from NIR spectroscopy has more unique sample information and can be separable. As well the scatter correction and standardization techniques extract information from the sample data of different origin and based on that information a discrimination of samples can be performed. Principal component analysis in this work for feature transformation and reduction provides number of uncorrelated data features that can be used for separability study and prediction analysis.

The separability index values calculated from different pre-processing techniques indicates that any simple clustering technique like k means clustering can discriminate the samples easily.

The prediction analysis using Principal component regression suggest that a liner discrimination and prediction model can be formed with the NIRS data with good accuracy and relatively minimal prediction error. Here also detrending is the most effective technique with minimum prediction error among all the pre-processing techniques.

So it can be concluded through this study that the techniques used in this research can effectively discriminate the Swertia chirayita samples collected from different geographical locations and a prediction model can also be formulated by PCR to predict the unknown samples of Swertia chirayita.

Further scope

In this study a separability and discrimination analysis was done on the NIRS data. The separability index value is good and the scatter plots also explains the same so if a simple clustering technique is applied on the data then it is going to provide a basis of discrimination upon which new data also can be classified and later on used for training.

The sample can be analysed by HPLC for percentage presence of phytochemicals discussed above for quantification and quality assessment of the present samples. So that a well calibrated statistical model can be formulated in future.

For statistical inference more number of samples can be analysed using the above technique that can give a better understanding about the structures of different samples of Swertia chirayita.