A Report

on

Prediction of heavy metal ions and its concentration from spectroscopy data

By

Pratyush Ramsharan Goel
ID No. 2017B5A70899P
Deepak Jain
ID No. 2017B5A30935P

Under the supervision of:

Dr Raj Kumar Gupta
Professor, Department of Physics

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE PILANI, PILANI CAMPUS

October 2021

Acknowledgements

During the course of this project, we had continuous support, encouragement and guidance from our supervisor, Dr Raj Kumar Gupta and mentor Ms Parul Taneja and their invaluable contribution to the report.

We are extremely grateful to Dr Rakesh Choubisa (HoD, Physics department) for encouragement and motivation for the project, ensuring us of every support possible from him and the department overall.

We are also thankful to all the Teaching staff of the Physics department for support in all the other ways, thus helping us in successfully completing the project work. We would also like to thank all the non-teaching staff for their contributions and support.

Abstract

Toxic heavy metal detection in water sources is crucial due to the detrimental social and environmental threats these metals pose. Traditional methods rely on expensive and sophisticated technologies, limiting their availability for on-site detection. This project demonstrates the application of machine learning algorithms to identify heavy metal ions (Cd, Co, Pb and Hg) and its concentration from a given water sample. The water sample is treated with H2TPP solution and its UV-VIS and X-ray fluorescence spectroscopy is observed to identify the unknown parameters.

Table of Contents

1. Introduction

Many developments in industry have brought plenty of new products and services. These advancements cause not only producing useful technology but also more usage of chemicals. The ecological contamination is responsible for destruction of environmental systems, increase in illnesses and diseases in humans, plants and animals. Three major types of ecological contamination are the following: atmospheric [1,2]; water [3,4] and soil pollution [5,6]. The atmospheric and water contamination can lead to huge disasters in a short time [7–11]. Many processes involve chemicals, which are harmful to the human health and the environment when they're released. Heavy metal ions are toxins due to their harmful impact to human health even at very low concentrations. They can be released easily into the environment (water, air, soil) via human activities and processing of natural resources. Thus, early detection of the heavy metal ions is a vital task in order to protect the environment and avoid pollution.

2. Literature Survey

Heavy metal ion detection remains a significant challenge due to the low concentrations of these ions, as well as the requirement for complex sampling processes and the use of expensive equipment. Inductively coupled plasma/atomic emission spectrometry (ICP-AES), inductively coupled plasma/mass spectrometry (ICP-MS), and atomic absorption spectroscopy (AAS) are few of the traditional methods for heavy metal ion studies. These analytical techniques (ICP-MS, ICP-AES, AAS or AFS) are highly sensitive and can selectively determine the concentrations of the different metal ions[12].

Several types of analytical methods have found applications in detecting and analysing heavy metal ions in water [13]. Traditional analytical methods for heavy metals detection provide benefits such as precision, sensitivity, versatility (suitable for a wide range of elements), and a high limit of detection. On the other hand, they have a number of drawbacks, including expensive instruments, difficult sample pretreatment, and complex analytical processing, all of which are incompatible with monitoring applications. Several different types of analytical procedures will be briefly demonstrated:

2.1. Atomic absorption spectroscopy

Atomic absorption spectroscopy (AAS) is a technique used extensively for heavy metal ion trace determination in all sample varieties. It is useful for one single element per analysis determination [14].

2.2. Inductively coupled plasma spectrometry

Inductively coupled plasma spectrometry (ICP) a multi-element analysis technique. It can be subdivided into two techniques: i.e. inductively coupled plasma-optical emission spectroscopy (ICP-OES) [15], and inductively coupled plasma-mass spectrometry (ICP-MS) [16].

2.3. Laser induced breakdown spectroscopy

Laser induced breakdown spectroscopy (LIBS) is a multipurpose technique. It accomplishes fast analysis of heavy metals in water and provides monitoring as well. Its unique spectral signatures identify the elements in water [17].

2.4. X-ray fluorescence

X-ray fluorescence (XRF) is an elemental analysis technique which is primarily used for excitation source, provided by X-rays tubes, which cause sample elements to emit X-ray photons of a characteristic wavelength and detectors are used to detect and analyse the secondary radiation (X-ray photons) [18].

2.5. Neutron activation analysis

Neutron activation analysis (NAA) is an extremely sensitive technique for determining different heavy metals concentrations put together on sensitivities, where concentration

of the elements is determined by reviewing the spectra of the radioactive sample emissions [19].

2.6. Ion chromatography

Ion chromatography (IC) is a simple technique for the concurrent analysis and quantification of heavy metals ions in the solution. It is an HPLC technique in which ion exchange resins are applied for simultaneous determination of many ionic species in aqueous solutions on ppm and ppb scale. This method proves advantageous over the spectroscopic methods for cation analysis [20].

2.7. UV-VIS spectrometry

UV-VIS spectrometry (UV-VIS) is founded on the concept of molecular absorption and proved useful in the determination of anions and cations concentrations at low scale, that's challenging to evaluate by the utilisation of AAS. There are many advantages of this method including its easy to use, simplicity, speedy and economical measurements of heavy metal ions in low to high concentrations [21].

3. Methodology

We focussed on using X-ray fluorescence and UV-VIS spectroscopy, interacting with 0.01 mM H2TPP solution, to collect 34 spectras each for all four heavy metal ions (Cd, Co, Pb and Hg) at varying concentrations ranging from 10^{-3} mM to 30 mM. These spectra are intensity vs wavelength plots, where the wavelength for UV-VIS spectroscopy ranged 250A^{o} - 800A^{o} (551 wavelengths) and for X-ray fluorescence ranged 450 nm - 800 nm (351 wavelengths). These wavelengths represent the dimensionality of the explanatory variable. The embedding vector stores the intensity of spectra and the length of the vector is 551 and 351 for UV and X-ray respectively.

Our goal is to identify the heavy metal ion and its concentration given a valid spectra. This is a trivial case of classification and regression. Classification is about predicting a label which in this case is heavy metal ions and regression is about predicting a quantity i.e., concentration. So, once the complete data was collected, it was split into 2 categories: train and test as shown in Fig 2-5 (Appendix section 6.1 and 6.2) to feed it to the ML model. The train-test split ratio for X-ray fluorescence and UV-VIS spectroscopy is 0.3 and 0.33 respectively.

3.1. Classification

Classification in machine learning and statistics, is a supervised learning approach in which the computer program learns from the data given to it and makes new observations or classifications. To satisfy our requirement, we used a Linear SVC (Support Vector Classifier) model [22]. The objective of a Linear SVC is to fit to the data and return a "best fit" hyperplane that divides, or categorizes the data. After getting the hyperplane, we can test the classifier on the test dataset and validate the "predicted" class with its actual labels. Here in our case we have only four labels/ classes (Cd, Hg, Co and Pb) and the hyperplane separates the classes in 551 and 351 dimensions for UV and X-ray respectively.

3.2. Multivariate Multiple Linear Regression

Multiple linear regression (MLR) [23], also known simply as multiple regression, is a supervised learning statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression is to model the linear relationship between the explanatory (independent) variables and response (dependent) variables.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon \quad [23]$$

where, for i=n observations and p = number of wavelength:

y_i = dependent variable (Concentration of heavy metal ion)

x_i = explanatory variables (intensity of spectra)

β_0 = y-intercept (constant term)

β_p = slope coefficients for each explanatory variable

ϵ = the model's error term (also known as the residuals)

As we have 4 classes / labels we define four linear regressions each corresponding to one heavy metal ion. Once we get the test labels from classification, we select the regression corresponding to that test label and predict the concentration w.r.t the spectra. Also, for X-ray fluorescence we normalise the value before passing to the regression model by taking log (base 10) of the intensity values as they are of the order ~ 10^7. This helps in achieving better correlated results.

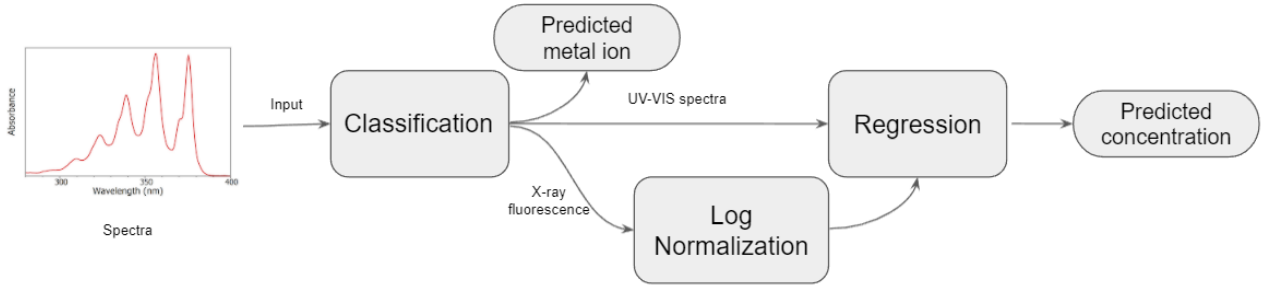

Fig 1: Flowchart of the complete ML model.

4. Results

- For identification of metals, we achieved a 100% accuracy using SVM classifier for both the datasets.
- As linear regression models also give negative values, we hard capped these values <= 0.001 mM to 0.001 mM.
- The average covariance score for UV-VIS and X-ray fluorescence is 0.9766 and 0.9522 which shows that the regression model is able to fit the test dataset.

Heavy Metal Ions	UV-VIS spectroscopy	X-ray fluorescence
Cd	0.9764	0.9612
Co	0.9977	0.9856
Hg	0.9833	0.9810
Pb	0.9493	0.8810
Avg Score	0.9766	0.9522

Table 1: Covariance score UV-VIS spectroscopy and X-ray fluorescence.

5. Discussion

Classification shows a very good accuracy but the regression is not able to show accurate predictions for concentration < 1 mM (refer appendix 6.3 Concentration Prediction Table, Table 2 and 3). To improve the model we tried many Machine learning and Deep learning techniques. We designed a Multi Layer Perceptron network (MLP) [24] with three hidden layer based and were able to reduce this error. Due to less dataset, the results from the neural network were not reproducible. For higher concentration of heavy metal ions, regression predicts the results very close to the original concentration.

6. Appendix

6.1. UV-VIS Spectroscopy Data

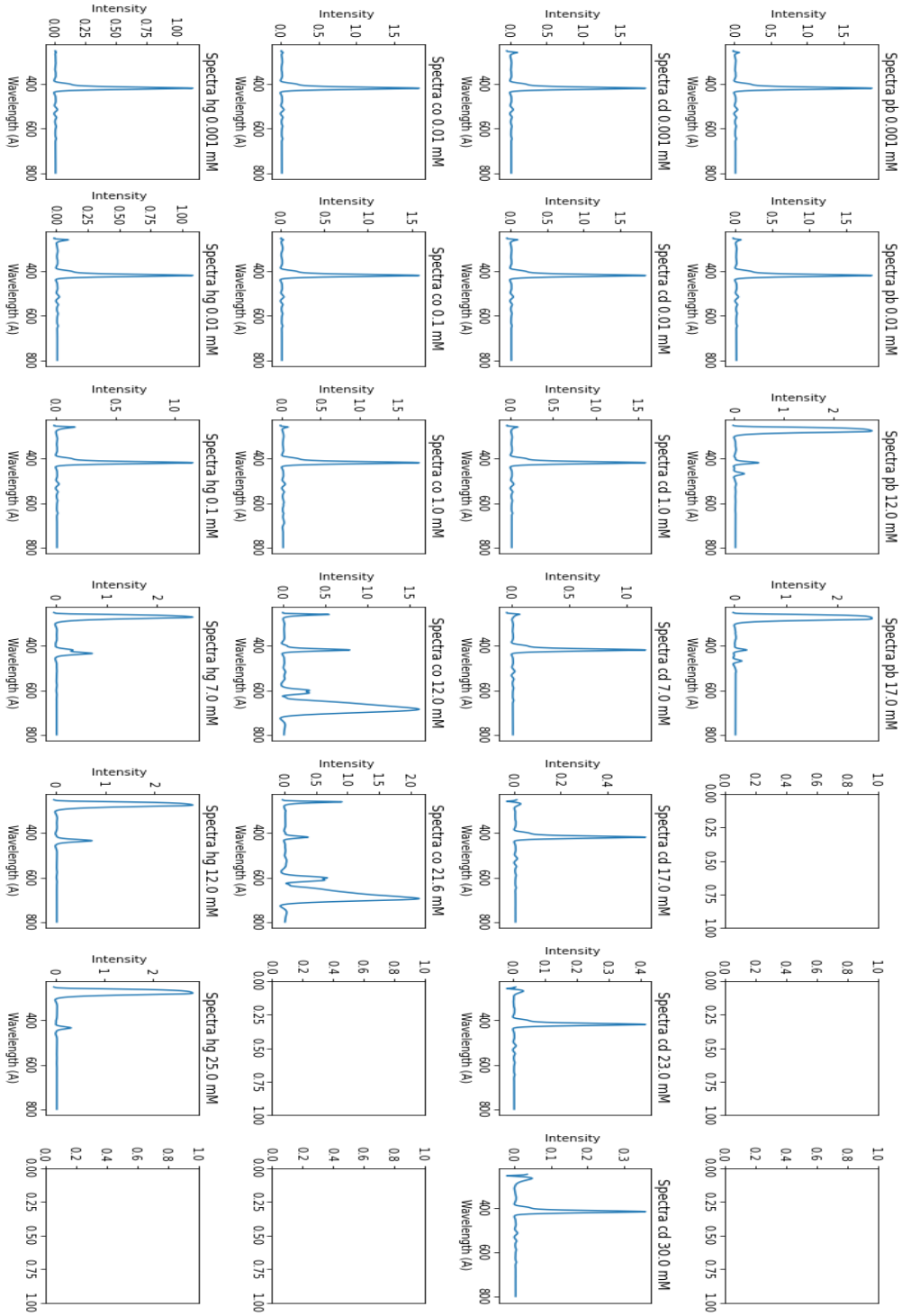

Fig 2: Train dataset.

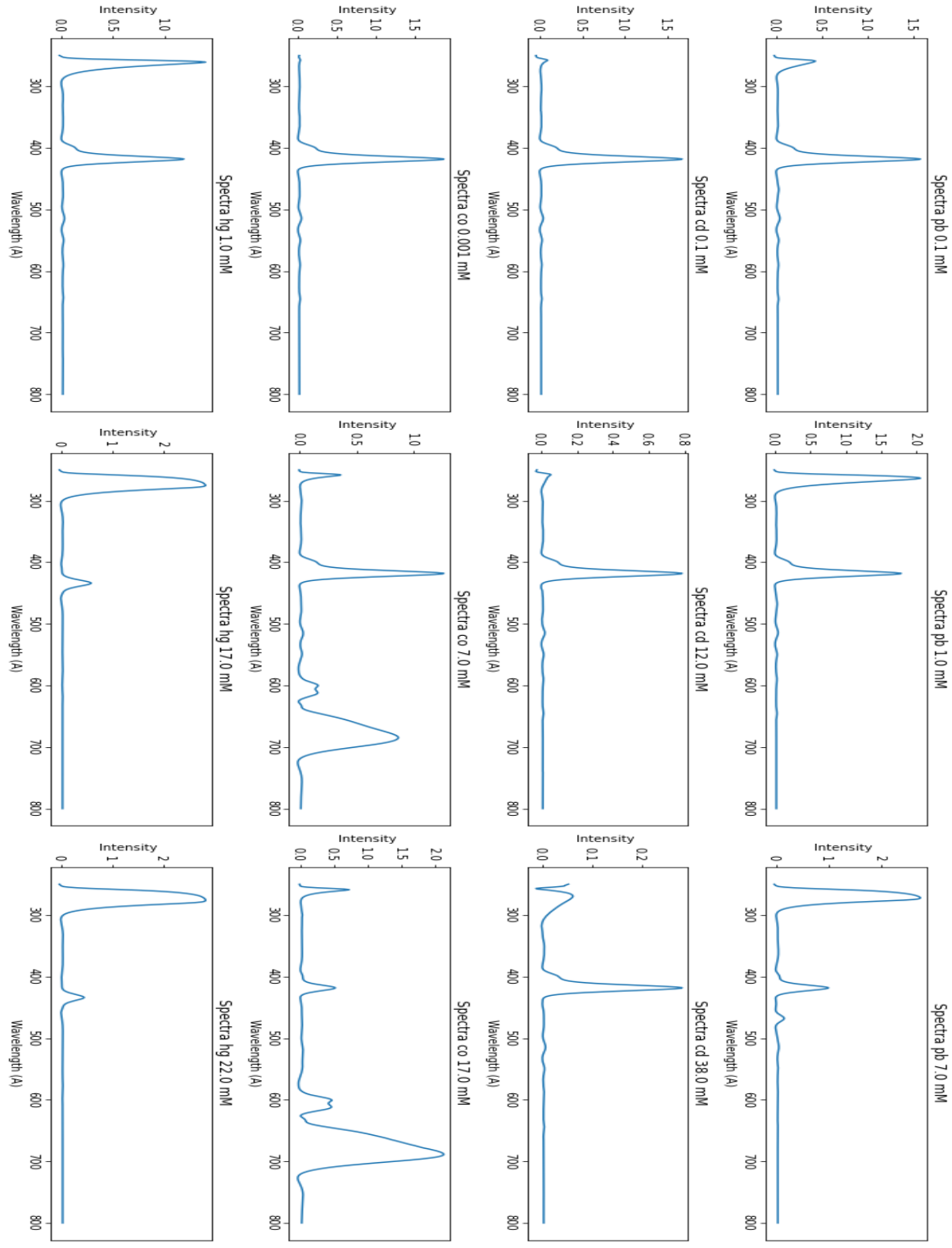

Fig 3: Test dataset.

6.2. X-ray fluorescence Data

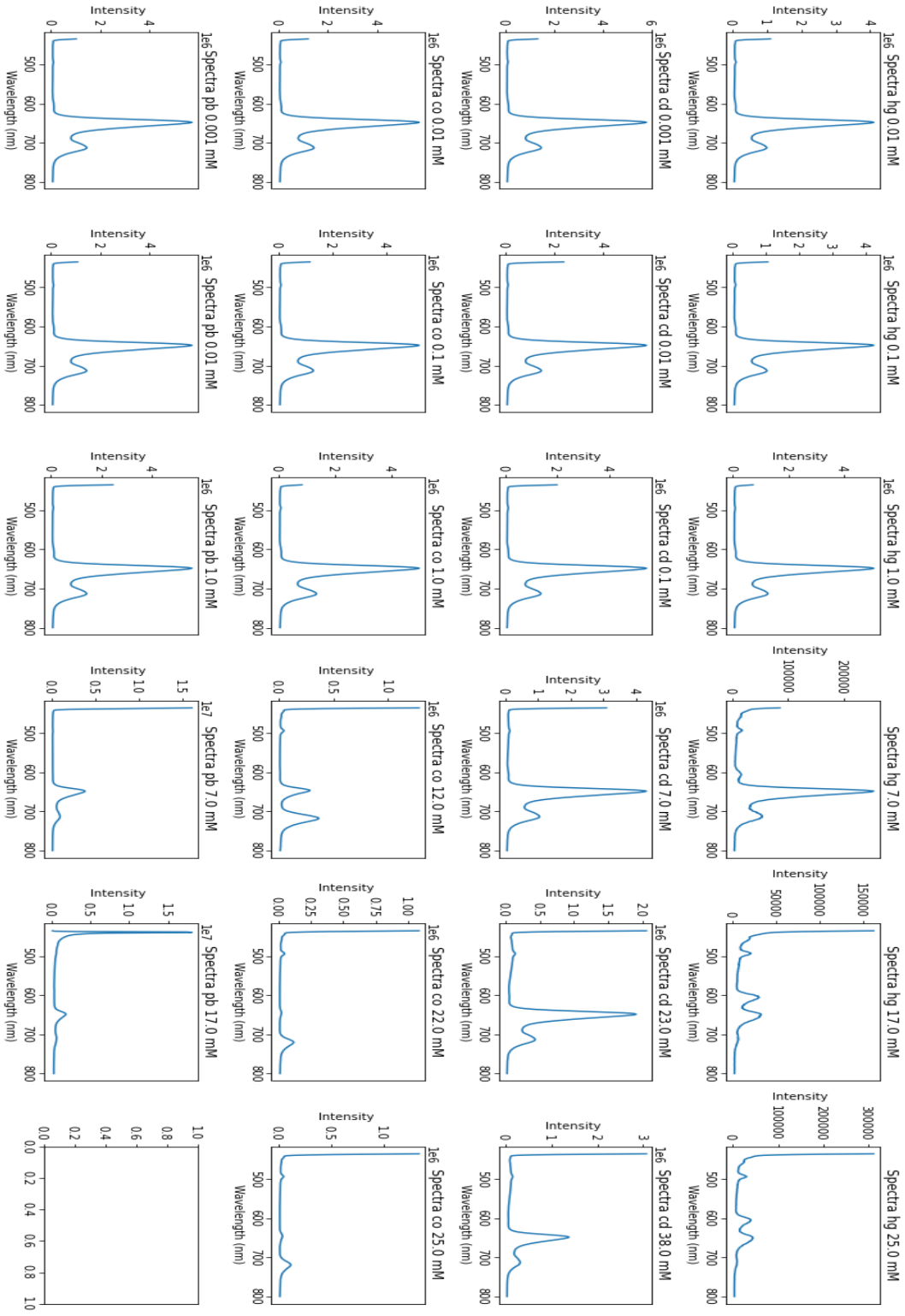

Fig 4: Train dataset.

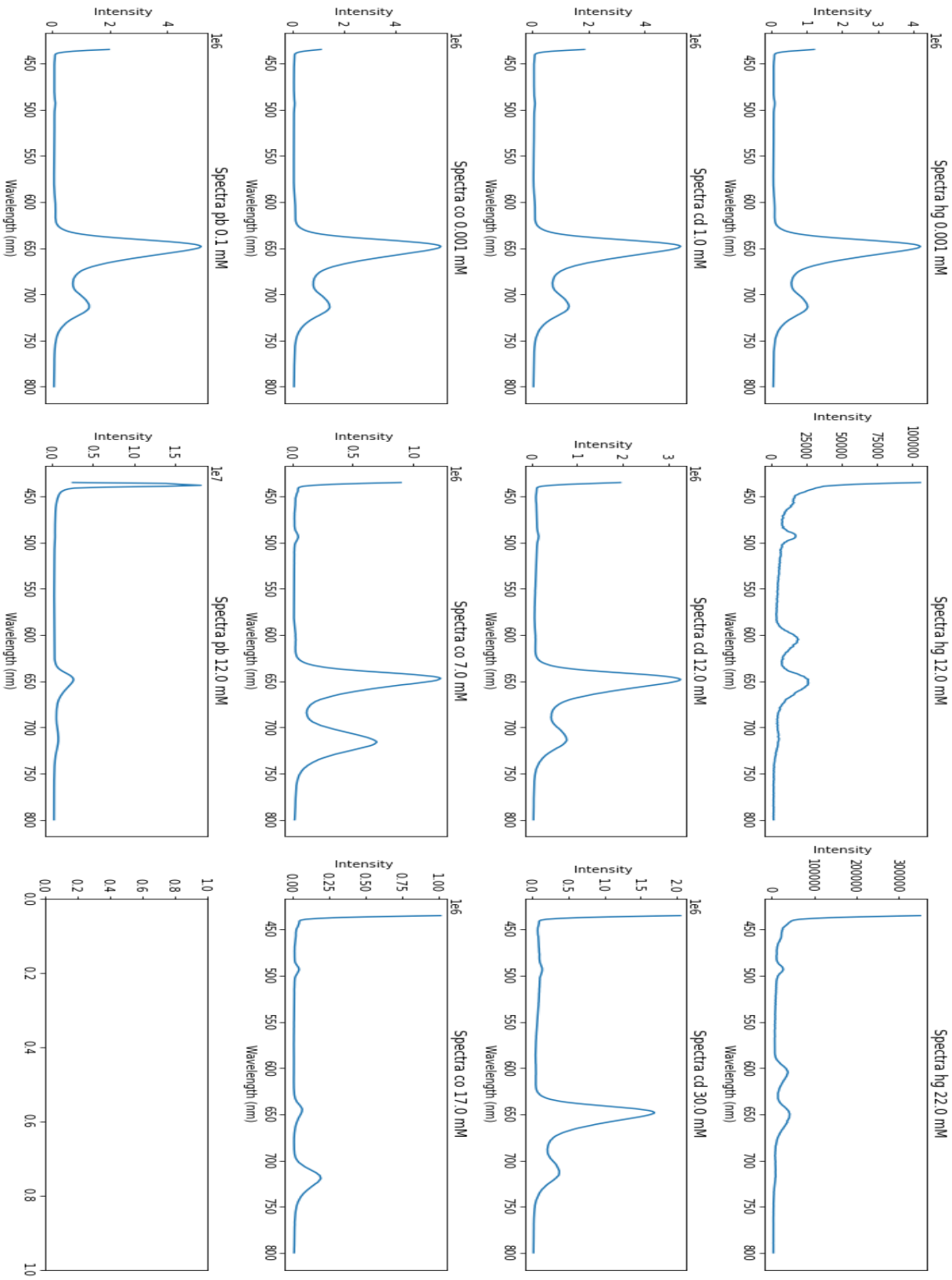

Fig 5: Test dataset.

6.3. Concentration Prediction Table

Heavy Metal Ion	Actual Concentration (in mM)	Predicted Concentration (in mM)
Cd	0.1	1.313
	12	13.946
	38	41.522
Co	0.001	0.001
	7	7.193
	17	16.459
Hg	1	2.837
	17	17.208
	22	21.236
Pb	0.1	0.265
	1	0.555
	7	5.904

Table 2: Predicted concentration for UV-VIS spectroscopy data with MAE = 0.994

Heavy Metal Ion	Actual Concentration (in mM)	Predicted Concentration (in mM)
Cd	1	0.001
	12	8.952
	30	27.759
Co	0.001	0.001
	7	6.034
	17	18.075
Hg	0.001	0.001
	12	13.741
	22	23.141
Pb	0.1	1.128
	12	14.714

Table 3: Predicted concentration for X-ray fluorescence data with MAE = 1.359

7. References

1. W.H. Smith, *Air Pollution and Forests Interactions between Air Contaminants and Forest Ecosystems* (Springer-Verlag, New York, 1981)
2. C.A.J. Appelo and D. Postma, *Geochemistry, Groundwater and Pollution* (CRC, Balkema Roterdam, 2005).
3. A. Olness, J. Environ. Qual. 24, 383–383 (1995) doi:10.2134/jeq1995.00472425002400020024x.
4. J.D. Sartor, G.B. Boyd and F.J. Agardy, J. Water. Pollut. Control. Fed. 46, 458 (1974).
5. P.C. Brookes, Biol. Fertil. Soils 19, 269 (1995). doi:10.1007/BF00336094.
6. K. Verschueren, *Handbook of Environmental Data on Organic Chemicals*, (John Wiley and Sons, Inc., New York, 2001), Vol. 1.
7. F.A. Gifford and S.R. Hanna, Atmos. Environ. 1967, 131 (1973). doi:10.1016/0004-6981(73)90202-3.
8. A.C. Stevenson and D.M. Elsom, Prog. Phys. Geogr. 18, 312–312 (1994). doi:10.1177/030913339401800216.
9. M. Moradi, L. Tulli, J. Nowakowski, M. Baljozovic, T. Jung and P. Shahgaldian, Angew. Chem. Int. Ed. 56, 14395 (2017). doi:10.1002/anie.201703825.
10. W.C. Tan, D. Qiu, B.L. Liam, T.P. Ng, S.H. Lee, S.F. van Eeden, Y. D'Yachkova and J.C. Hogg, Am. J. Respir. Crit. Care. Med. 161, 1213 (2000). doi:10.1164/ajrccm.161.4.9904084.
11. J.R. Stedman, Atmos. Environ. 38, 1087 (2004). doi:10.1016/j.atmosenv.2003.11.011.
12. L. Eddaif, A. Shaban & J. Telegdi (2019) Sensitive detection of heavy metals ions based on the calixarene derivatives-modified piezoelectric resonators: a review, International Journal of Environmental Analytical Chemistry, 99:9, 824-853, DOI: 10.1080/03067319.2019.1616708
13. S. Jignesh, K. Vineeta, S. Abhay and K. Vilasrao, Int. J. Res. Pharm. Chem. 2 (1), 146 (2012).

14. R.A. Mohamed, A.M. Abdel-Lateef, H.H. Mahmoud and A.I. Helal, Chem. Speciat. Bioavailab. 24 (1), 31 (2012). doi:10.3184/095422912X13257005726800.
15. O.V.S. Raju, P.M.N. Prasad, V. Varalakshmi and Y.V.R. Reddy, Int. J. Innov. Res. Sci. Eng. Technol. 3 (2), 9743 (2014).
16. M. Batsala, B. Chandu, B. Sakala, S. Nama and S. Domatoti, Int. J. Res. Pharm. Chem. 2 (3), 671 (2012).
17. F. Zhao, Z. Chen, F. Zhang, R. Li and J. Zhou, Anal. Methods. 2 (4), 408 (2010). doi:10.1039/b9ay00160c.
18. L. Borgese, A. Zacco, E. Bontempi, M. Pellegatta, L. Vigna, L. Patrini, L. Riboldi, F.M. Rubino and L.E. Depero, J. Pharm. Biomed. Anal. 52 (5), 787 (2010). doi:10.1016/j.jpba.2010.02.030.
19. V.P. Guinn and D. Wagner, Anal. Chem. 32 (3), 317 (1960). doi:10.1021/ac60159a005.
20. C. Sarzanini and M.C. Bruzzoniti, Trends Analyt. Chem. 20 (6–7), 304 (2001). doi:10.1016/S0165-9936(01)00071-1.
21. C.O.B. Okoye, A.M. Chukwuneke, N.R. Ekere and J.N. Ihedioha, Int. J. Phys. Sci. 8 (3), 98 (2013). doi:10.5897/IJPS12.670.
22. Cortes, C., & Vapnik, V. (1995). *Support-vector networks. Machine Learning*, 20(3), 273–297.
23. Jobson J.D. (1991) *Multiple Linear Regression. In: Applied Multivariate Data Analysis. Springer Texts in Statistics. Springer, New York, NY.* https://doi.org/10.1007/978-1-4612-0955-3_4
24. Haykin, S. (1994). Neural networks: a comprehensive foundation. Prentice Hall PTR.