Proceedings of the 8th World Congress on Civil, Structural, and Environmental Engineering (CSEE'23)
Lisbon, Portugal – March 29 – 31, 2023
Paper No. ICEPTP 192
DOI: 10.11159/iceptp23.192

Quantification of Hydrocarbon Contamination in Soil Using Hyperspectral Data and Deep Learning

Rafic Ayass[1], Samir Mustapha[1], Darine Salam[2]
[1]Mechanical Engineering, American University of Beirut
[2]Civil and Environmental Engineering, American University of Beirut
Bliss Street, Beirut, Lebanon
rsa98@mail.aub.edu; sm154@aub.edu.lb; ds40@aub.edu.lb

Abstract - Petroleum and its products undergo large scale production, transportation, and storage which make them prone to spills and leakages into the environment. Petroleum contamination in terrestrial environments, particularly soil bodies, is common and holds major consequences on food and crops, microbial communities, the atmosphere, the water sphere, public health and safety, and the soil itself, and therefore, requires immediate detection and assessment in case contamination is present. In this paper, advanced machine learning methods are used to predict petroleum hydrocarbon contamination in soil using hyperspectral image data. Hyperspectral imaging combines imaging and spectroscopy and can detect petroleum hydrocarbons using the characteristic absorption features in hydrocarbon reflectance spectra. Laboratory prepared soil samples are contaminated with crude oil and scanned with a hyperspectral camera in a laboratory setup. The data collected is used to train deep learning models to predict, quantitatively, the amount of petroleum present in soil samples. To make predictions, a first model is built to use spectral data from a single pixel while a second model is built to use spectral and spatial data by using two adjacent pixel spectra as input. The results show good performance for both models, with the two-pixel model achieving a better mean square error of 0.48 on a testing dataset compared to the mean square error of 0.628 for the single pixel model on the same testing data. Therefore, hyperspectral imaging contains valid spectral and spatial information that are beneficial for assessing petroleum contamination in soil with good accuracy.

***Keywords*:** hyperspectral imaging, oil spills, soil pollution, petroleum hydrocarbons, deep learning, artificial intelligence, neural networks

1. Introduction

Petroleum remains a hugely important resource worldwide, with applications in many industries, mainly transportation, and amounts to trillions of barrels being produced annually. However, its large-scale transportation and storage can lead to accidents such as spills that can have devastating effects on ecosystems and economies that linger for years [1]. Even on the short term, spills are flammable and contain toxic chemicals that are harmful upon exposure. The process of detecting and responding to oil spills is critical and must be done quickly and accurately to prevent further damage.

Petroleum Hydrocarbons (PHCs) are a complex mixture of thousands of different chemicals along a wide range of types, properties, and behaviors that are extracted from underground reservoirs as crude oil and taken to refineries to be separated into different products based on their uses [2]. Petroleum releases into the environment originate from crude oil or crude oil-based products such as gasoline and diesel fuel. Furthermore, petroleum spills in marine and terrestrial environments eventually make their way to soil bodies or marine sediments [1]. However, the sources of petroleum hydrocarbon contamination in soil could also be petrogenic such as industrial emissions, pyrogenic such as burning of fossil fuels, or biogenic due to living organisms [3]. In addition, soil petroleum contamination holds major consequences on the soil itself, food and crops, microbial communities, the atmosphere, the water sphere, and public health and safety [4].

Many methods have been developed to measure the levels of petroleum contamination in soil both quantitatively and qualitatively such gas chromatography and fluorescence spectroscopy. Most of the common methods suffer in terms of portability and analysis time and require sample extraction and preparation which involves the use of hazardous solvents. On the other hand, Visible-Near Infrared (Vis-NIR) spectroscopy is proven to be a portable, quick, and non-invasive method for hydrocarbon assessment in soil with good accuracy [5]. In the Vis-NIR spectroscopy range (800-2500 nm), hydrocarbon spectra originate mainly from combinations or overtones of C-H stretching modes of saturated CH_2 and terminal CH_3 or aromatic C-H functional groups resulting in absorption at 1200, 1725, and 2310 nm [5, 6]. However, spectroscopy alone fails to give information on spatial distribution of the spectra obtained which limits its application [7]. On the other hand, hyperspectral imaging combines spectral and spatial features by integrating spectroscopy and imaging in a sensor that acquires several dozens or hundreds of images of contiguous wavelengths, assigning every pixel in the hyperspectral image its own spectrum over a contiguous wavelength range represented by an array of spectral bands. The resulting hyperspectral image is a three-dimensional data cube composed of two spatial dimensions (x rows and y columns) as well as one spectral dimension (λ wavelengths).

Due to the direct relation between PHCs and their reflectance spectra, the spectra can be used to make predictions on the levels of PHC contamination in soil mediums quantitatively and qualitatively by analyzing it [5, 6]. Several different methods have been used to analyze spectral data to assess petroleum hydrocarbons in soils. Linear methods for spectral data analysis include regression analysis such as linear regression and logistic regression, multivariate techniques, Partial Least Squares (PLS) and Principal Component Analysis (PCA) [8-10]. However, factors affecting hyperspectral remote sensing such as spectral variability due to varying image conditions and mixed pixel spectra due to contributions from different materials introduce complexities that require nonlinear analysis methods [11]. Nonlinear methods for spectral data analysis include neural networks and kernel-based transformations [6, 12].

This paper's objective is to build predictive models to estimate the amount of petroleum hydrocarbon contaminant in contaminated soils using hyperspectral images while making use of spatial and spectral features through deep neural networks. The remainder of this paper is structured as follows: a methodology section that describes three main processes which are petroleum contamination of soil samples, hyperspectral data collection, and deep learning model training, a results section to report and analyse the results, and a conclusions section to discuss the significance of our work.

2. Methodology

The experimental method begins with preparing samples of PHC contaminated soil. For the PHC, this research uses Arabian light crude oil as the PHC of choice being the major export-grade oil in the Arab region and pure Ottawa sand (2 to 0.05 mm grain size) as the soil type of choice, mostly representing beach sediment. The hyperspectral camera used is the Hyspex SWIR-384 with a 930-2500 nm spectral range, 288 spectral channels. A laboratory setup was constructed to operate the camera and scan samples in a controlled environment which shown in Fig. 1. The setup consists of a light source to keep the samples under constant illumination using halogen bulbs that emit a constant stream of light within the camera's spectral range, a blackout fabric cover that prevents external light from interfering with the measured spectra, and a custom open containment box to place the samples as they are scanned which is painted black to prevent light scatter. A Spectralon cube is placed near each sample during a scan which acts as a calibration target to obtain the relative reflectance of a sample by dividing the irradiance of the sample by the irradiance of the cube [13].

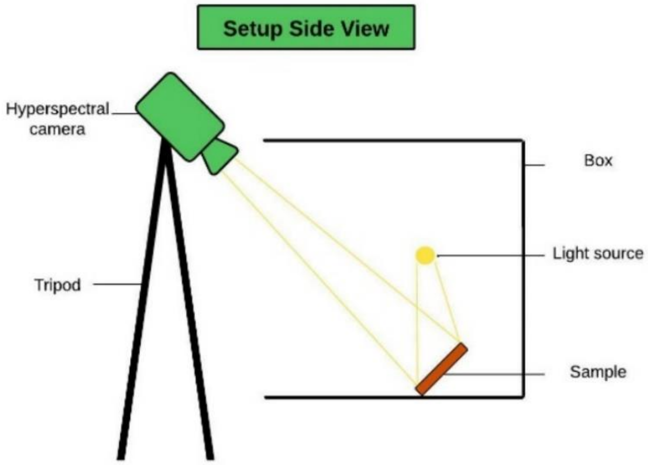

Fig. 1: Side view of the laboratory setup.

Contaminated soil samples were prepared through incremental addition of crude oil to 50 grams of Ottawa sand in a petri dish. Initially, the uncontaminated sample was scanned with the camera. An aliquot of 0.5 mL of crude oil was then added to the sand, thoroughly mixed to ensure even distribution of the contaminant in the sample and scanned again with the camera. The process was repeated with 0.5 mL of crude oil being added each time to the same sand sample until a total of 2 mL of oil were added. After that, crude oil addition was performed with increments of 1 mL, with sample scanning taking place after each increment. This process was repeated until a total volume of 13 mL of crude oil were added to achieve the sample saturation.

As for the deep learning model, two models were trained based on two input types. The first model takes as input the spectrum of a single pixel, i.e., the 288 spectral bands of that pixel. The architecture of this model consists of four hidden layers with 30 neurons each and an output layer with one neuron corresponding to the predicted volume of oil in the sample. All neurons in all layers use the rectified linear activation function. The second model takes as input the spectra of two spatially neighbouring pixels that are consecutive in the x-dimension with the same y-dimension from the same hyperspectral image. Therefore, this model uses spatial as well as spectral features by taking as input the 288 spectral bands of two pixels in a local region which sums up to a total of 576 input values. The architecture of this model consists of four hidden layers with 60, 30, 20, and 10 neurons respectively and an output layer with one neuron corresponding to the predicted volume of oil in the sample. All neurons in all layers use the rectified linear activation function as well.

To train the models, pixel reflectance spectral data from the hyperspectral image of each increment were used. Only pixels inside the region of the petri dishes, i.e., sample pixels, were considered. To obtain the training dataset of the single-pixel model, the spectral bands of each pixel inside the sample were extracted, converted to reflectance, and assigned a label equal to the ground truth total volume of oil added to the sample when the scan was done. The dataset was split into 75% of pixels for training and 25% for validation. A testing set was formed by preparing new samples using the same pure sand and crude oil and contaminating 50 grams of sand at four random contamination levels, namely 0.2, 2.5, 7.5, and 10.5 mL of crude oil in each sample. The reflectance spectral bands of the pixels from the test samples scans and their labels, i.e., the ground truth volume of oil in the sample were used to form the testing dataset. Using the same data, the datasets of the second model were formed by extracting the reflectance spectral bands of every possible combination of two x-dimension consecutive pixels inside the same sample and assigning them a label equal to the ground truth volume of oil in the sample.

3. Results

The average spectra of each incremental sample, i.e., each oil contamination level, is plotted in Fig. 2 along with their standard deviations. Fig. 2 also shows the hyperspectral image cube that was taken of the sand sample before any petroleum contaminant was added. In the graph of Fig. 2, the x-axis corresponds to the wavelengths and the y-axis corresponds to the relative reflectance values limited between 0 for no light reflectance and 1 for maximum light reflectance at a given wavelength. The average reflectance spectrum shifts towards less light reflectance the more oil is added to the sample as shown in Fig. 2; i.e., the more oil volume in a sample, the lower the reflectance spectrum is in the plot. The PHC absorption feature dips for the samples with oil are visible at the 1725 and 2310 nm wavelengths with the help of the black vertical lines in the graph of Fig. 2. Moreover, the standard deviations in Fig. 2 indicate little intersection between the reflectance spectra of different contamination levels which allows us to assume that each contamination level has a unique range of reflectance spectra associated with it.

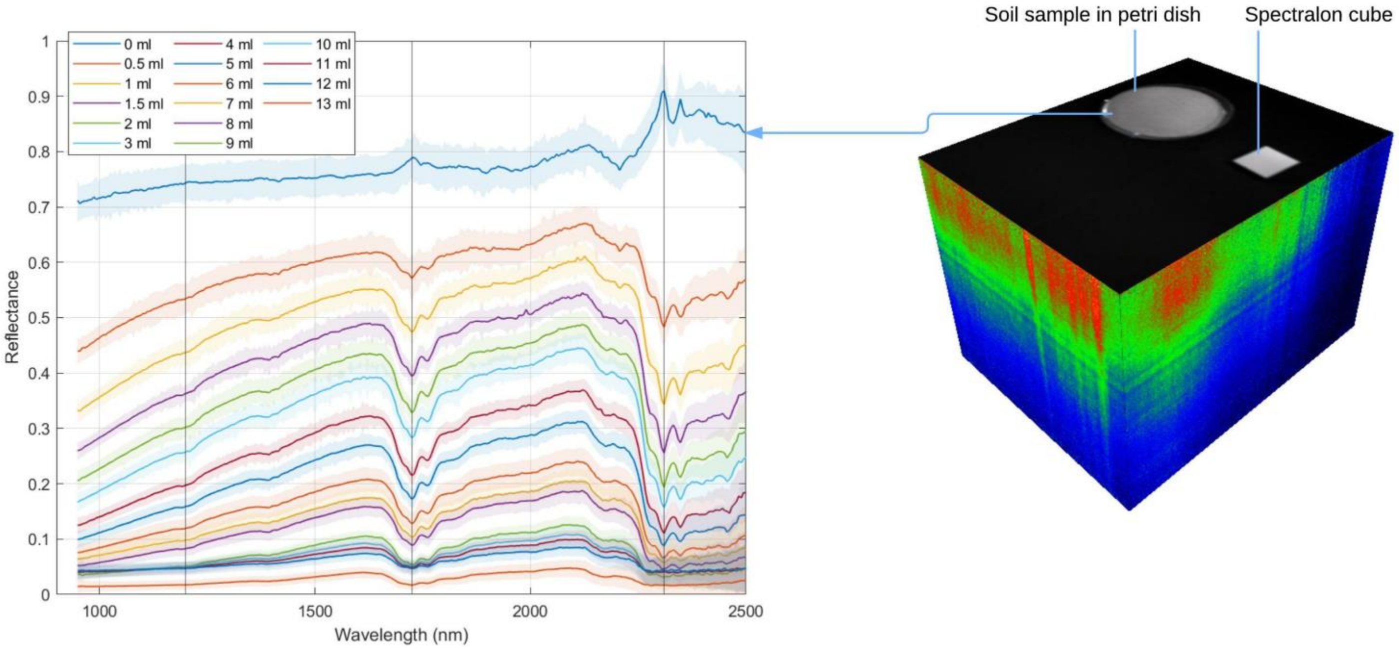

Fig. 2: Average reflectance spectra of the Arabian light oil and pure sand samples at different contamination levels to the left and an example hyperspectral image of the soil sample with 0 ml of contaminant to the right

To better visualize the absorption features, a normalization method called continuum removal is used which essentially transforms spectra to a common baseline for comparison [9]. Continuum removal is performed on the average reflectance spectra of each sample around the 1725 nm feature wavelength and is shown in Fig. 3. The depth and width of the absorption feature is directly proportional to the volume of oil in the sample, i.e., the deeper and wider the dip in the absorption feature, the higher the amount of oil in the sample. Therefore, we can assume that correlations between the reflectance spectra and the level of crude oil contamination in a sand sample exist and can be learned using a machine learning model.

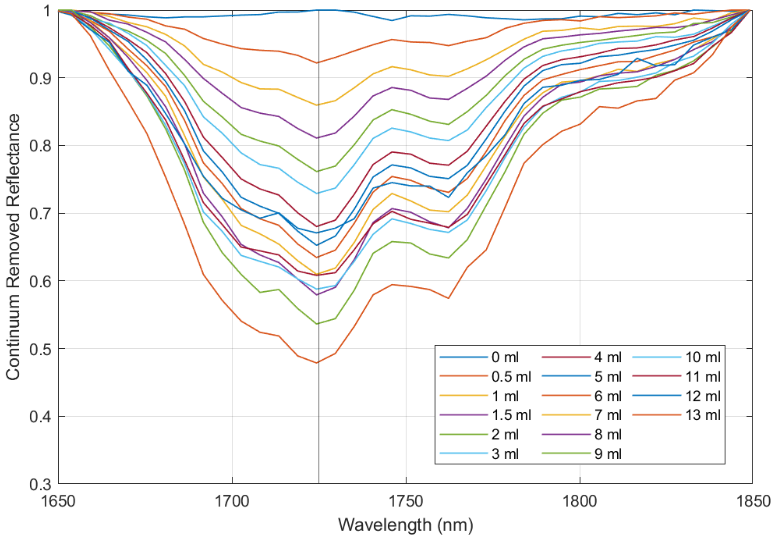

Fig. 3: Continuum removed reflectance spectra of the contaminated sand samples between 1650 and 1850 nm

From the training datasets formed, the deep learning models were constructed and trained for each model individually. The Mean Square Error (MSE) is a common metric used for regression problems in machine learning that uses the average squared difference between observed and predicted values over all instances in a dataset. The MSE was calculated using predictions from each model on the training, validation, and testing datasets and is shown in Table 1.

Table 1: Mean square errors of the two models on the datasets.

Model	**Training error**	**Validation error**	**Testing error**
Single-pixel model	0.093	0.119	0.628
Two-pixel model	0.101	0.128	0.48

The results of both the single-pixel and two-pixel models show good performance on the collected data. Low training and validation errors indicate that the models fit the training data well. The two-pixel model has slightly higher training and validation errors than the single-pixel model since it is easier to fit the single-pixel model to its data as it must adjust to less input values, 288 input bands, compared to the two-pixel model with 576 input bands. For both the single-pixel and two-pixel model, the largest errors are on the testing datasets. The reason is that the test data was created from unseen samples that had been contaminated separately to determine the ability of the models to generalize on new samples and scans. Therefore, it is expected to find a drop in performance while moving from training and validation data that are used to build and fit the model, to the testing data that is used to test the model's performance. Nonetheless, the testing data errors indicate good model performance, which can be improved further with more tuning of model hyperparameters. The results also indicate the benefit of adding spatial data along with the spectral one as input. Using spectral input from two locally adjacent pixels in a neighbouring region resulted in a two-pixel model that performs significantly better than its single-pixel counterpart on the testing data with 24% improvement in performance with respect to mean square error.

4. Conclusions

This study demonstrates the ability of hyperspectral data and deep learning techniques to accurately predict levels of petroleum contamination. Predictions of the single-pixel and two-pixel model on an independent testing dataset showed good performance and potential. A 24% performance increase was obtained with the two-pixel model over the single-pixel model. This highlights the importance of spatial data in hyperspectral imaging for making better predictions on PHC levels in soil. This method holds potential in aiding emergency response procedures by monitoring and detecting petroleum spills and contaminations in a quick and practical manner before further testing with analytical and laboratory techniques can be done. For the future, we aim to continue this research by creating samples from various other soil types and petroleum hydrocarbons to create more diverse data for model training. We also intend to explore further the effects of spatial data by training more sophisticated spectral-spatial models with different types of input from hyperspectral image data.

Acknowledgements

The author would like to acknowledge the financial support from the Middle East Partnership Initiative-Tomorrow's Leaders Graduate program at the American University of Beirut provided to Mr. Rafic Al Ayass with the following award number: SNEAAC19CA0037.

References

[1] S. Mishra, G. Chauhan, S. Verma, and U. Singh, "The emergence of nanotechnology in mitigating petroleum oil spills," *Marine Pollution Bulletin,* vol. 178, p. 113609, 2022/05/01/ 2022, doi: 10.1016/j.marpolbul.2022.113609.

[2] ITRC. "TPH fundamentals." Interstate Technology and Regulatory Council. https://tphrisk-1.itrcweb.org/4-tph-fundamentals/ (accessed.

[3] Sakshi, S. K. Singh, and A. K. Haritash, "Polycyclic aromatic hydrocarbons: soil pollution and remediation," *International Journal of Environmental Science and Technology,* vol. 16, no. 10, pp. 6489-6512, 2019/05/25 2019, doi: 10.1007/s13762-019-02414-3.

[4] S. Wang, Y. Xu, Z. Lin, J. Zhang, N. Norbu, and W. Liu, "The harm of petroleum-polluted soil and its remediation research," presented at the AIP Conference Proceedings, 2017.

[5] R. N. Okparanma and A. M. Mouazen, "Determination of Total Petroleum Hydrocarbon (TPH) and Polycyclic Aromatic Hydrocarbon (PAH) in Soils: A Review of Spectroscopic and Nonspectroscopic Techniques," *Applied Spectroscopy Reviews*, vol. 48, no. 6, pp. 458-486, 2013/08 2013, doi: 10.1080/05704928.2012.736048.

[6] A. M. Ahmed, O. Duran, Y. Zweiri, and M. Smith, "Quantification of Hydrocarbon Abundance in Soils Using Deep Learning with Dropout and Hyperspectral Data," *Remote Sensing*, vol. 11, no. 16, p. 1938, 2019/08/19 2019, doi: 10.3390/rs11161938.

[7] D. Wu and D.-W. Sun, "Advanced applications of hyperspectral imaging technology for food quality and safety analysis and assessment: A review — Part I: Fundamentals," *Innovative Food Science & Emerging Technologies*, vol. 19, pp. 1-14, 2013/07/01/ 2013, doi: 10.1016/j.ifset.2013.04.014.

[8] S. Chakraborty *et al.*, "Rapid Identification of Oil-Contaminated Soils Using Visible Near-Infrared Diffuse Reflectance Spectroscopy," *Journal of Environmental Quality*, vol. 39, no. 4, pp. 1378-1387, 2010/07/01 2010, doi: 10.2134/jeq2010.0183.

[9] R. D. P. M. Scafutto and C. R. d. Souza Filho, "Quantitative characterization of crude oils and fuels in mineral substrates using reflectance spectroscopy: Implications for remote sensing," *International Journal of Applied Earth Observation and Geoinformation*, vol. 50, pp. 221-242, 2016/08 2016, doi: 10.1016/j.jag.2016.03.017.

[10] J. G. Bray, R. A. V. Rossel, and A. B. McBratney, "Diagnostic Screening of Urban Soil Contaminants Using Diffuse Reflectance Spectroscopy," in *Proximal Soil Sensing*, R. A. Viscarra Rossel, A. B. McBratney, and B. Minasny Eds. Dordrecht: Springer Netherlands, 2010, pp. 191-199.

[11] D. G. Manolakis, R. B. Lockwood, and T. W. Cooley, *Hyperspectral Imaging Remote Sensing: Physics, Sensors, and Algorithms*. Cambridge: Cambridge University Press, 2016.

[12] S. Kang, G. Xiurui, T. Hairong, and Z. Yong-chao, "A New Target Detection Method Using Nonlinear PCA for Hyperspectral Imagery," *Bulletin of Surveying and Mapping*, p. 105, 2015.

[13] H. Zhang, Y. Yang, W. Jin, C. Liu, and W. Hsu, "Effects of Spectralon absorption on reflectance spectra of typical planetary surface analog materials," *Opt. Express*, vol. 22, no. 18, pp. 21280-21291, 2014/09/08 2014, doi: 10.1364/OE.22.021280.