CNN Approach for IR Spectral Classification (May 2020)

ABSTRACT Infrared (IR) is a section of electromagnetic radiation that is nearly invisible to the human eye. The IR spectrum of an object can be plotted by shooting the IR wavelengths at that object and then plotting the reflectance values. Research has been conducted in this area to see if there are defining characteristics of an IR spectrum that distinguish it from other IR spectra. Traditionally, spectral similarity algorithms like the Pearson Correlation Coefficient and Euclidean Distance have been used to measure the similarity among spectra. This approach comes with a shortcoming, however. Samples of the same compound can have large variances among reflectance values and can cause inaccuracies with the algorithms. An improved approach would be to analyze the overall trends of the spectrum and not focus on intensities. It is due to this shortcoming that we explore a convolutional neural network (CNN) approach to measuring the spectral similarity. This multifaceted approach explores classification based on spectral class and probabilistic matching.

INDEX TERMS infrared spectrum, convolutional neural network, spectral similarity

I. INTRODUCTION

Infrared (IR) spectroscopy is the study of spectra in the IR range. This research deals with a subsection of IR, the near-IR and mid-IR ranges. These ranges consist of around 0.3 to 15 micrometers. Due to the length of the wavelengths, this is nearly invisible to the human eye. An IR spectrometer is able to measure the reflectance at each wavelength by shooting those wavelengths at the object. The points can then be plotted into a graph and visualized like the figure below.

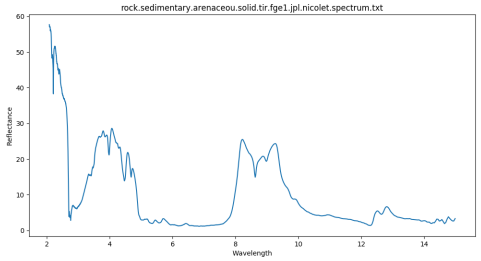


Figure 1. Sedimentary Rock IR Spectrum

IR spectroscopy is a powerful method that has been applied to many different fields like, honey validation [1], meat adulteration detection [2], and fish ageing [3]. This method offers a lot of promise because it is relatively cheap when being compared to other methods. If it is possible to create an effective approach to analyzing this field then it will offer many more financial opportunities. This research is dedicated to finding an approach that increases the effectiveness of this method.

The chemical composition of the object determines the resulting spectrum, thus making it a possibility to distinguish between different compounds. From Figures 2 and 3 below, you can see how a soil spectrum differs from a chloride spectrum.

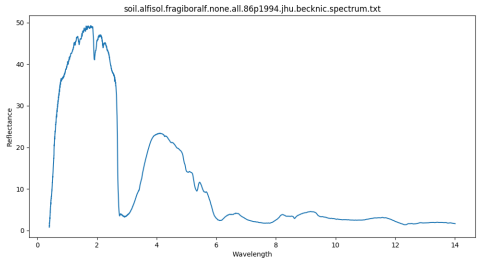


Figure 2. Soil IR Spectrum

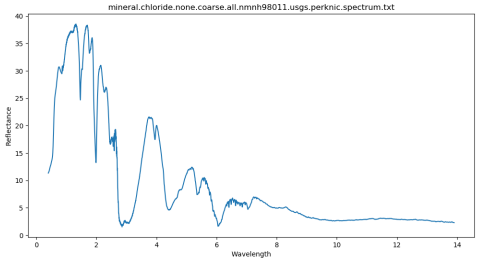


Figure 3. Chloride IR Spectrum

The distinction between compound groups is not trivial to make, however. IR spectroscopy is often less accurate than similar methods and reflectance can vary between samples of the same compound. The variance in reflectance often depends on the chemical complexity of the compound. Figures 4 and 5 show two samples of vegetation that are very chemically similar. Figures 6 and 7 show two samples of metamorphic rock that have a high variance among reflectance values.

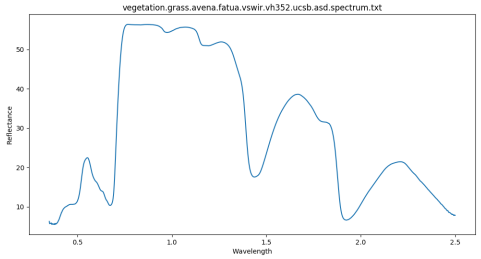


Figure 4. Sample 1 of a vegetation IR spectrum (low variance)

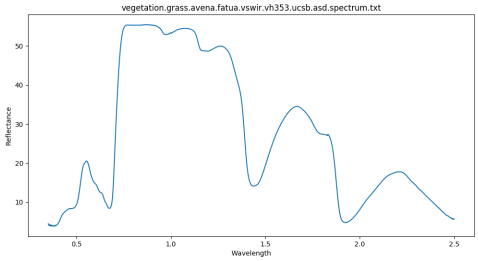
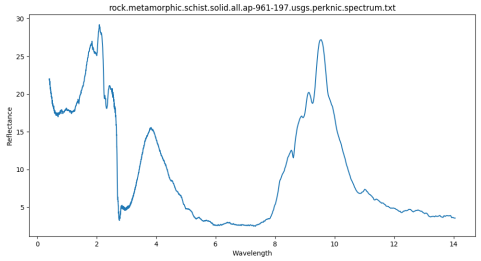


Figure 5. Sample 2 of a vegetation IR spectrum (low variance)



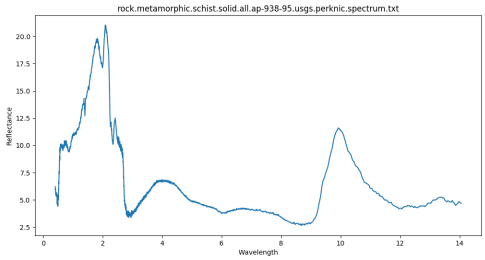


Figure 6. Sample 1 of metamorphic rock IR spectrum (high variance) Figure 7. Sample 2 of metamorphic rock IR spectrum (high variance)

The difference between these two samples highlights one of the many difficulties of analyzing spectra. Traditional methods have been proven effective to classify samples with low variance. The issue arises however when we try to analyze samples with high variance. In many real world settings, the variance is going to vary greatly and a method that can be effective at both is highly desired.

Previous research has been conducted on the ECOSTRESS dataset. The dataset consists of IR samples of various compounds. Each compound can be broken up by its spectral type, class, subclass, and particle size. Spectral type is the most general classification. Spectral class is the next level in specificity and subclass being the next after that. Particle size is used to differentiate since there are varying particle sizes among the subclass grouping. The goal of the research project was to create an effective way to classify spectra using four similarity algorithms, Euclidean Distance, Manhattan Distance, Cosine Similarity, and Pearson Correlation Coefficient. The approach was to group iteratively by the four groupings previously mentioned. Thus, each grouping contains samples that have the same

classifications for each level. Then a pair of spectra are randomly selected out of each grouping. A spectrum is then queried against every other spectrum in the sample dataset. The effectiveness is based on how well the calculated best match is to the queried spectrum's type, class, subclass, and particle size. Meaning if the spectral type is the same for the calculated best match and queried spectrum then it is considered correctly classified at the type level. This happens for the other classifications as well. Out of the four algorithms used, the Pearson Correlation Coefficient performed the best. The figure below shows the accuracy achieved.

Accuracy (%)	Type	Class	Subclass	Particle Size (Best Match)
Pearson Correlation Coefficient	93.58	73.83	49.51	30.37

Figure 8. Accuracy of Pearson Correlation Coefficient on ECOSTRESS dataset.

As you can see, the accuracy is not very high when trying to find the best possible match. This showed some of the issues with the similarity algorithms and left a need for a more accurate method. Due to these shortcomings in spectral comparison and analysis, we propose a CNN approach in this multifaceted research project. The first area of investigation focuses on extracting compound defining trends and classifying a spectrum based on those trends. The second area of investigation seeks to emulate a similarity algorithm. Similarity algorithms compare two spectra and output a score denoting the likeness of the two spectra. The model proposed in this section seeks to produce a probability that two spectra are the same or not.

II. Literature Review

Use Case 1: Validating Honey

An area that has used FTIR has been utilized is validating the authenticity of honey [1]. Many areas around the globe have a huge problem with the demand for honey outweighing the supply. This issue has made many companies and producers resort to adding various ingredients to increase the supply. Ingredients like corn syrup, sugar, and water can be added to honey and make it very hard to distinguish between the real and fake honey. Visually it is almost impossible to tell, but the chemical composition has changed. This allows the use of IR spectroscopy to be used on the honey. The IR spectrum can then be analyzed in four wavelength ranges to validate whether the honey is fake or not.

Use Case 2: Meat Adulteration

The Food and Drug Administration (FDA) has incorporated FTIR into its meat quality check [2]. Some companies add other types of meats into their beef in an attempt to lower production costs and increase profits. This raises concerns because some of those meats might be against religious beliefs or considered taboo in certain parts of the world. Adulterating beef is also dishonest to customers that believe they are purchasing only beef. The FDA has used many different techniques to minimize these adulterations. Many of these techniques require high tech lab equipment and are very invasive as many of them require testing on the DNA and protein levels. FTIR, on the other hand, is not as invasive and can provide results much faster than other techniques. Due to the agileness of the method, massive quantities can be tested.

Use Case 3: Ages of Fish

The government uses FTIR to measure fish ages in federally managed waters [3]. Surveying fish ages in a body of water is used as a way of measuring population health. The aging assessment is crucial to measuring the stock price for the fishing industry and dictates what the government needs to do to maintain the species. There are several aging processes that involve analyzing the fish otolith, ear stones. These methods, however, are inaccurate. Most labs can only guarantee the precision of their estimate being within 10-20% of the actual age. There is a way to use FTIR on the fish otolith. Otolith is composed mostly of calcium so they can measure the absorbance of a certain protein to determine the age.

III. Problem Statement

Traditionally, similarity algorithms have been used to classify spectra. This is done by taking an unknown spectrum and using an algorithm to compute the similarity with known compounds among a dataset. The closest known compound is then considered most spectrally similar to the unknown spectrum. Some common similarity algorithms are Euclidean Distance, Manhattan Distance, Cosine Similarity, and Pearson Correlation Coefficient [4, 5].

Euclidean Distance: $\sqrt{\sum_i (x_i - y_i)^2}$ Manhattan Distance: $\sum_i |x_i - y_i|$

Euclidean and Manhattan distance behave very similarly when computing the similarity between two spectra. The essence of the similarity score is based on the distance between the two points at each wavelength. In the case of these algorithms, the larger the score, the less similar they are. The algorithms measure the distance between two spectra so a distance of zero indicates the same spectrum. The difficulty in this approach is highlighted from the point made with Figures 6 & 7, the reflectance can vary at individual wavelengths. This causes the algorithms to measure differences even though the trends are similar.

Cosine Similarity: $\frac{A \bullet B}{||A||\,||B||}$ Pearson Correlation Coefficient: $\frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}}$

Cosine Similarity and Pearson Correlation Coefficient have been used to measure trends in spectra [4, 5]. The issue, although somewhat mitigated, is also due to the variance in reflectance values. The algorithms weight the peaks and valleys of the spectrum by the intensity of them. Which like Manhattan and Euclidean distance, causes the algorithms to measure differences even if the trends are similar.

From the issues previously mentioned, similarity algorithms struggle with analyzing overall trends. Due to this, we use a CNN approach to extract these distinguishing features.

IV. Proposed Approach

Classification

The proposed approach for classifying a spectrum by its spectral class begins with the ECOSTRESS dataset. Due to the broadness of spectral type, many samples can be under the same grouping that is not chemically similar. It is due to this that we move to the next level, spectral class. The process begins by grouping each spectrum by its class. We then use the data to train and validate the model.

Probabilistic Matching

The proposed approach for creating a probability model starts with the ECOSTRESS dataset. The data is grouped similar to the previous research project with the similarity algorithms. The dataset is grouped by type, class, subclass, then particle size. To generate the data for CNN, we begin by creating entries for two classes, matching, and non-matching. Matching entries will consist of a spectra pair from the same grouping. Non-matching entries will consist of a spectra pair that are from different groupings. The data is then used to train and validate the model.

IV. Proposed Solution

We propose a CNN implementation for the problem stated. To approach this problem we create two separate models to handle each section of the problem. An IR spectrum of a compound will have many peaks and valleys within a given range. The difficulty when classifying an IR spectrum is that the spectrum varies between samples. Since there is such a high variance between samples, it is better to look at the overall trends of the spectra. This makes CNNs well suited for spectral classification. Through features like dimensionality reduction, CNNs are well equipped to identify the key characteristics that define an IR spectral group.

Classification

The optimal performing CNN was a two-dimensional model. The activation function used at every layer, except for the output layer using softmax, was Leaky ReLU with an alpha value of 0.01. Leaky ReLU on average gave more accurate results than similar activation functions like ReLU. The input layer is a convolutional layer with 64 nodes. The input is then followed by a 2x2 max-pooling layer. Further into the hidden layers, another convolutional and max-pooling layer exist with the same specifications. The model is then flattened and run through a dense layer of 128 nodes. The output layer shrinks down to the number of classes, in this case, 42, and a softmax activation function is applied. The loss method is categorical cross-entropy with adam as the optimizer. Several dropout layers were added in between layers to combat overfitting.

Probabilistic Matching

The optimal performing CNN was a two-dimensional model. The activation function used at every layer, except for the output layer using sigmoid, was Leaky ReLU with an alpha value of 0.01. Leaky ReLU on average gave more accurate results than similar activation functions like ReLU. The input layer is a convolutional layer with 16 nodes. The input is then followed by a 2x2 max-pooling layer. Further into the hidden layers, is two sets of convolutional and max-pooling layers. The model is then flattened and run through a dense layer of 32 nodes. The output layer is then minimized to one node and a sigmoid activation function is applied. The loss method is categorical cross-entropy with adam as the optimizer. Several dropout layers were added in between layers to combat overfitting.

V. Implementation Details

Sample Setup

Both fields of research involve the same preprocessing to every sample. Each sample needs to be converted into a three-dimensional version of itself. Height and width represent the first two dimensions and the third is represented by the color channel, RGB. By plotting the points and placing markers in one of the color channels, we represent the spectrum.

Classification

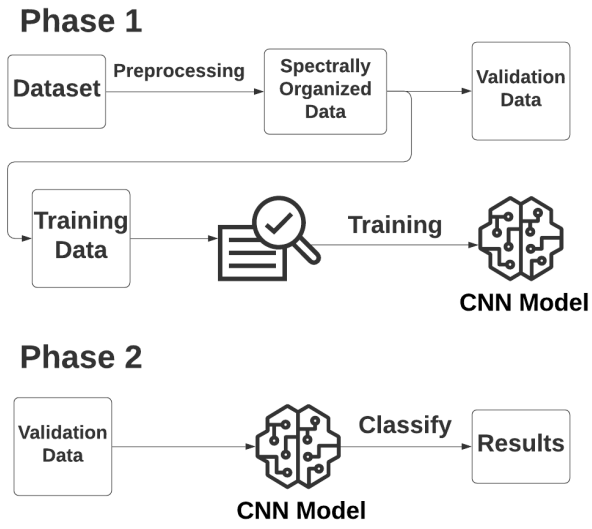


Figure 9. Flowchart for the classification process

The figure above features the major checkpoints of the Classification process. The process starts by preprocessing the dataset. We start by grouping every sample into its spectral class. The data is then split into training and validation data. The training data is fed into the CNN with specifications described in the Proposed Solution section. We can then use the trained model to classify the validation data.

Probabilistic Matching

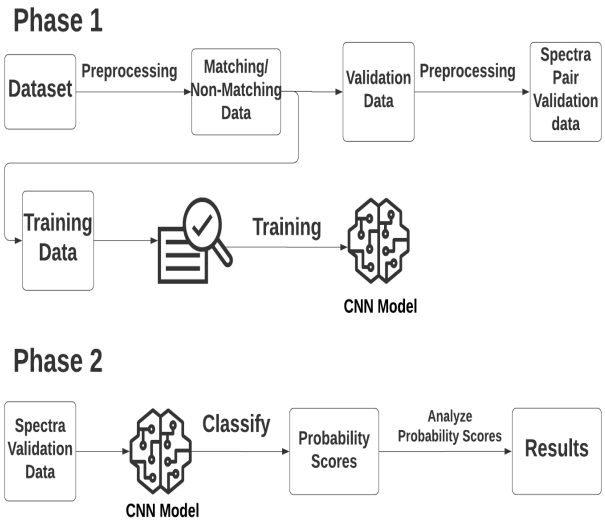


Figure 10. Flowchart for probabilistic matching

The figure above features the major checkpoints of the Probabilistic Matching process. The preprocessing begins by grouping up each spectrum by its type, class, subclass, then particle size. The data is then split into training and validation data. The training data gets further processed into matching and non-matching entries. For matching entries, iteratively go through each group and combine each spectrum sample with every other spectrum sample. An example of this is if we have samples A, B, and C in a group. The result would be the spectra pairs of AB, AC, and BC. To create the pairs, create a new three-dimension version of the two spectra and place them into separate color channels. The result will be two color channels holding a spectrum and the third channel being empty. We create on-matching entries by randomly selecting two spectra that are not of the same spectral type. This helps to ensure that the two samples will have different trends. Validation data is also processed by selecting two random spectra from each group. Every possible spectra pair is then created for all of the selections. The training data can now be used to train the model. Once the model is created, use the validation data to retrieve the probability scores. Similar to the previous research involving the similarity algorithms, get the accuracy for each spectrum at the type, class, subclass, and particle size levels.

VI. Evaluation

Classification

To measure the effectiveness of this approach, we apply the accuracy metric. We consider a successful classification if a spectrum is correctly classified as its spectral class. Classification of any other type is considered a failed classification. Based on this metric, the optimal model achieves an accuracy of 86.2% when classifying by spectral class. A similar study that utilized ECOSTRESS dataset used all similarity algorithms previously mentioned. With the Pearson Correlation Coefficient scoring the highest, similarity algorithms only reached accuracies of ~74%.

Probabilistic Matching

To measure the effectiveness of this approach, we apply the accuracy metric. We consider a successful matching if the queried spectrum is matched with its closest possible match. Any other match that is not the most spectrally similar sample is considered a failure.

Accuracy (%)	Pearson Correlation Coefficient	CNN
Type	93.58	86.88

Class	73.83	60.00
Subclass	49.51	43.13
Best Match	30.37	37.50

Figure 11. Probabilistic Matching and Pearson Correlation Coefficient Accuracies

In the top three levels, the Pearson Correlation Coefficient outperforms the CNN approach. The last level, however, the CNN outperforms the similarity algorithm by ~7%.

VII. Conclusion and Discussion.

In this article, we highlighted a novel approach that utilizes CNNs. Based on the results section, the CNN approach is more effective in terms of classifying spectra. The optimal model scored 12% higher than the best similarity algorithm used in previous research. The CNN was able to extract the overlying trends much more effectively than the similarity algorithms and will offer a promising area of research in the future. The Probabilistic Matching approach is a little more nuanced. The Pearson Correlation Coefficient was better at correctly classifying spectra in the type, class, and subclass categories, but was outperformed when selecting the best possible match. Due to the best match accuracy being higher it still offers promise for future research. Future work will involve testing on different datasets to test the robustness of the approach.

VIII. References

1. M. Sahlan, S. Karwita, M. Gozan, H. Hermansyah, Y. Masafumi, Y. Young Je, and D. K. Pratami, “Identification and classification of honey's authenticity by attenuated total reflectance Fourier-transform infrared spectroscopy and chemometric method,” Veterinary World, vol. 12, no. 8, pp. 1304–1310, Aug. 2019.
2. E. Deniz, E. G. Altuntaş, N. İğci, B. Ayhan, D. Ö. Demiralp, and K. Candoğan, “DETECTION OF PORK, HORSE OR DONKEY MEAT ADULTERATION IN BEEF-BASED FORMULATIONS BY FOURIER TRANSFORM INFRARED SPECTROSCOPY.,” GIDA / The Journal of FOOD, vol. 45, no. 2, pp. 369–379, 2020.
3. T. E. Helser, I. Benson, J. Erickson, J. Healy, C. Kastelle, and J. A. Short, “A transformative approach to ageing fish otoliths using Fourier transform near infrared spectroscopy: a case study of eastern Bering Sea walleye pollock (Gadus chalcogrammus).,” Canadian Journal of Fisheries & Aquatic Sciences., vol. 76, no. 5, pp. 780–789, May 2019.
4. J. Li, D. Hilbert, S. Fuller, Numerical methods for comparing fresh and weathered oils by their FTIR spectra, Analyst 132 (2007) 792-800.
5. J. Li, D. Hilbert, S. Fuller, G. Vaughn, A comparative study of point-to-point algorithms for spectra, Chemom. Intell. Lab. Syst. 82 (2006) 50-58.