# GeoAI Competition: The Hyperspectral Imaging Challenge



**Date:** 12 November 2023

**Team Name:**

MTTDATA

**Team Members:**

Brock Bennett

Isabella Lieberman

John Ramirez

Nathan Zlomke

# Table of Contents

# Executive Summary

ESA Φ-lab, in collaboration with KP Labs and their partner QZ Solutions, has launched an extraordinary initiative to transform agriculture's future by leveraging in-orbit processing. The primary goal is to enhance farm sustainability by utilizing the latest advancements in Earth observation and artificial intelligence. This approach helps address the challenge of affordable food production and contributes to environmentally friendly agriculture practices.

One of the critical aspects is providing farmers with timely information about soil parameters to optimize their fertilization processes. This optimization can lead to selecting more suitable fertilizer mixes and reducing overall fertilizer usage. Currently, the traditional method for quantifying soil parameters is labor-intensive and time-consuming. It involves collecting soil samples in the field and sending them to specialize labs for chemical analysis. Moreover, the limited number of sampling points in the field, often spread across large areas, needs to be improved to ensure the accuracy of test results. In-situ analysis could be more scalable and more time-inefficient.
The proposed solution is to harness innovative airborne and satellite hyperspectral imaging technology to promote more sustainable agriculture practices, contributing to a better future for our planet.

# Company Overview

AI for Good is organized by ITU in partnership with 40 UN Sister Agencies. The goal of AI for Good is to identify practical applications of AI to advance the United Nations Sustainable Development Goals and scale those solutions for global impact. It is the leading action-oriented, global & inclusive United Nations platform on AI.

# Problem Statement

The objective of this challenge is to advance the state-of-the-art in soil parameter retrieval from hyperspectral data in preparation for the forthcoming Intuition-1 mission. In March 2021, a campaign was conducted over agricultural regions in Poland, involving extensive ground samplings coordinated with airborne hyperspectral measurements from sensors mounted on an

aircraft. The hyperspectral data consists of 150 contiguous hyperspectral bands spanning from 462 to 942 nm, with a spectral resolution of 3.2 nm, matching the spectral range of the hyperspectral imaging sensor aboard Intuition-1.

Intuition-1 is a 6U-class satellite mission developed by KP Labs, designed to observe Earth using a hyperspectral instrument and an onboard computing unit equipped for in-orbit data processing using artificial intelligence. It will be the world's first satellite with the processing power required for advanced analysis of hyperspectral images in orbit.

In this challenge, the primary aim is to automatically estimate specific soil parameters, including potassium (K), phosphorus pentoxide (P2O5), magnesium (Mg), and pH.

## An Overview of Hyperspectral Imaging

*Summary of Hyperspectral Imaging*

Hyperspectral imaging is a technique that collects and processes information across the electromagnetic spectrum to obtain the spectrum for each pixel in an image. It allows for identifying objects and materials by analyzing their unique spectral signatures. Hyperspectral imaging applications include food quality & safety, waste sorting and recycling, and control and monitoring in pharmaceutical production.

Spectral imaging is imaging that uses multiple bands across the electromagnetic spectrum. While the RGB camera uses three visible light bands (red, green, and blue) to create images, hyperspectral imagery makes it possible to examine how objects interact with many more bands, ranging from 250 nm to 15,000 nm and thermal infrared. The study of light-matter interaction is called spectroscopy or spectral sensing.

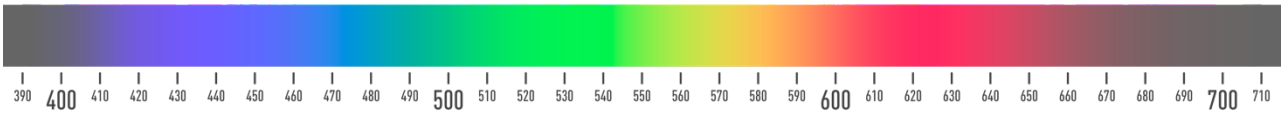


*Figure 1*

Hyperspectral imaging involves using an imaging spectrometer, also called a hyperspectral camera, to collect spectral information. A hyperspectral camera captures a scene’s light,

separated into individual wavelengths or spectral bands. It provides a two-dimensional image of a scene while simultaneously recording the spectral information of each pixel in the image. The result is a hyperspectral image, where each pixel represents a unique spectrum. This unique spectrum can be compared to fingerprints. Since every material and compound reacts with light differently, their spectral signatures also differ. Just like fingerprints can be used to identify a person, the spectra can identify and quantify the materials in the scene.

## Background Research

*Background on Hyperspectral Imaging*

According to Specim, "Hyperspectral imaging is an increasingly used technique in industry, research, and remote sensing. The data provided by hyperspectral imaging systems can be used during inspection to locate, sort, or quantify the concentration of various materials that are invisible to common cameras or the human eye."

GlobalNewswire notes, "Global hyperspectral imaging markets will reach US $49.4 billion through 2030 from US $22 billion in 2022." As part of a quickly growing industry, hyperspectral imaging is at the forefront of many who value what is beneath them. Hyperspectral imaging makes it efficient, and the funds for the hyperspectral imaging market are increasingly rising every year, so correctly processing hyperspectral images effectively and efficiently is a priority to optimize the money being thrown at it.

GlobalNewswire also mentions that "Many hyperspectral imaging system developers and manufacturers are tech-driven and lack vision for the scalability of the business and lack skills for marketing." Those system developers and manufacturers only speak their jargon and care more about the crunch to optimize hyperspectral imaging. Still, they need to pay more attention to communication between others who could be more familiar. The system developers and manufacturers need to prioritize breaking down the hyperspectral imaging for others unfamiliar with understanding it and help them utilize it for scalability and marketability.

*Specificity of Agriculture*

To gain excellent specificity of agriculture, it needs to find the soil properties in a non-destructive way while maintaining a high spatial resolution. With this all-in mind, the outcome would lead to farmers having more of an effective way to apply fertilizers and other products to help supplement healthy soil. All while ensuring costs are low and have slim to no hazardous environmental impact.

*Phosphorous in Soil*

Phosphorus is an essential nutrient for plant growth. Phosphorous deficiency can lead to stunted growth, reduced crop yields, and poor-quality produce. Excessive phosphorous in soil can lead to environmental concerns such as water pollution and algae blooms that result in adverse effects in water bodies.

*Potassium in Soil*

Potassium is also an essential nutrient for plant growth. Potassium is vital in plants' photosynthesis, water uptake, and disease resistance. A potassium deficiency can lead to reduced crop yields, lower quality produce, and increased susceptibility to pests and diseases. Potassium levels in soil can influence soil pH and nutrient uptake. An imbalance in potassium can affect plants' availability and absorption of other essential nutrients.

*Magnesium in Soil*

As a third essential nutrient for plant growth, magnesium is a vital component of the chlorophyll molecule and is crucial for photosynthesis. An imbalance in magnesium levels can affect the uptake of other essential nutrients. Magnesium regulates soil pH and can act as a liming agent to help raise soil pH in acidic soils. Adequate magnesium in soil can improve a plant’s ability to withstand stress, such as drought and extreme temperatures. Magnesium also helps prevent nutrient imbalances that could lead to nutrient runoff and water pollution.

# Data Exploration

The distributions of each soil parameter were plotted and are presented below. The values underwent Yeo-Johnson power series transformations to achieve higher model performance by normalizing the data.

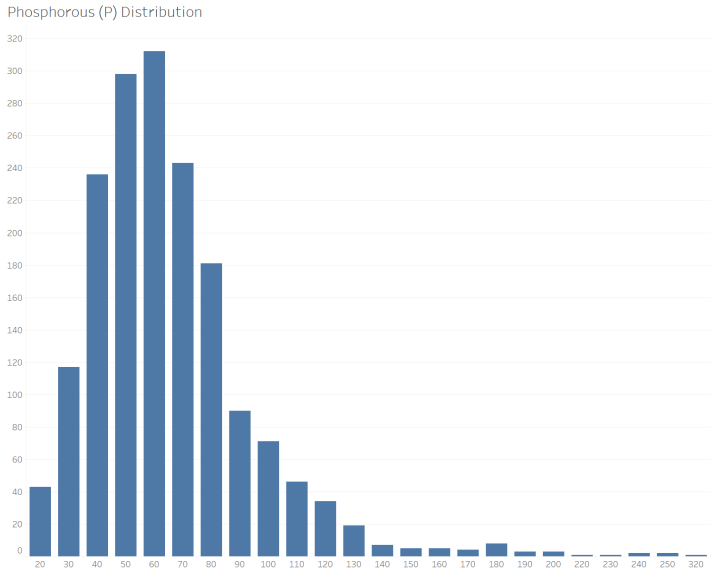


*Figure 2*

Phosphorous (P) peak, or mode, is located at 60 ppm with right skew.

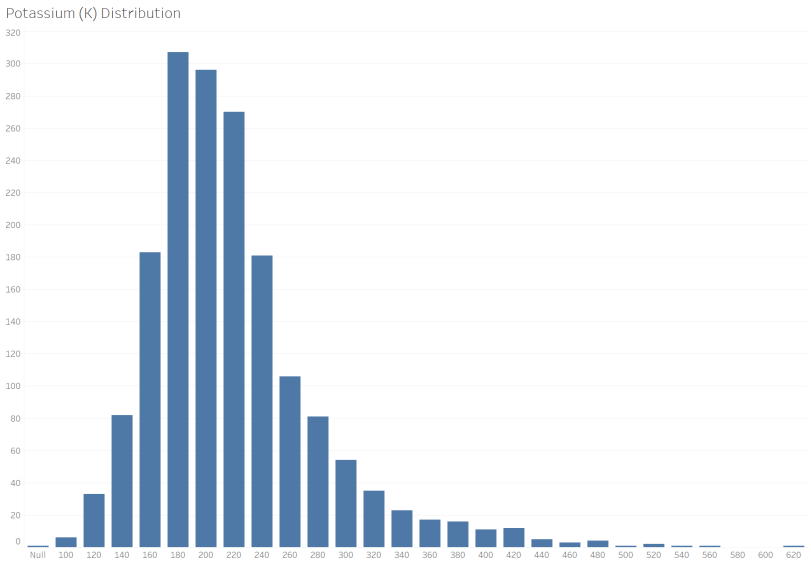


*Figure 3*

Potassium (K) highest frequency is located at 180 ppm. Distribution is right skewed.

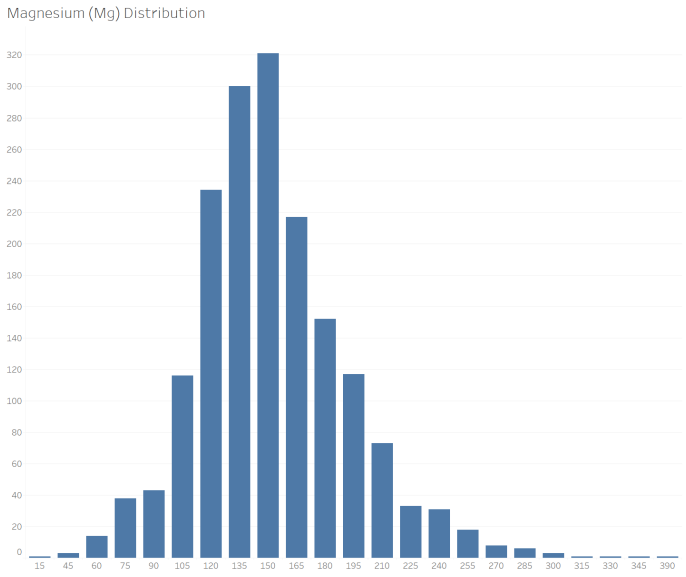


*Figure 4*

Magnesium has bell-shaped distribution with mode at 150 ppm.

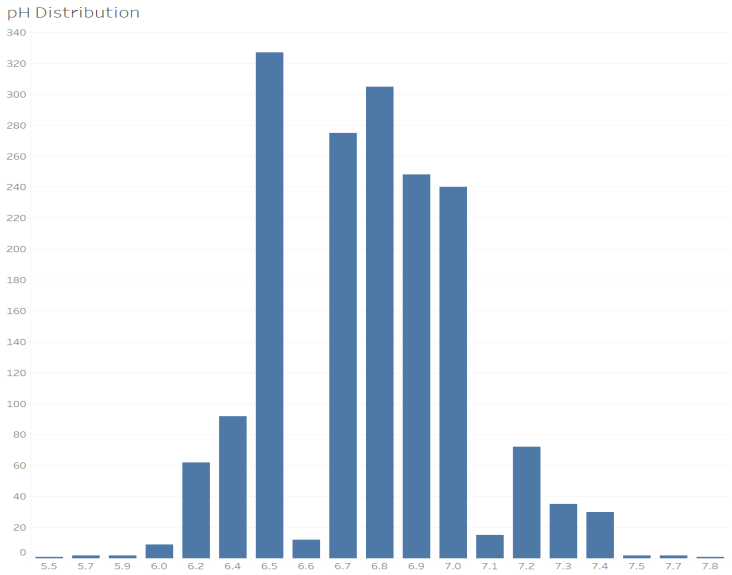


*Figure 5*

The most frequent reading for pH was 6.5. Shape is mostly normal, although large dips are observed.

# Data Cleaning and Reduction

*Transforming the Data*

During data exploration, it was discovered that the distributions of some soil parameters were skewed, as seen in Figure 1. Since these are target variables, it is not recommended to perform power series transformations; however, yeo-johnson transformations were performed on predictor variables. It can help achieve better accuracy in predicting the target variables.

*Scaling the Data*

Both predictor and target variables were scaled before initializing the model. It ensures all variables in the model are on the same scale, and those with higher magnitudes do not bias the model. It can also lessen the impact of outlier features.

# Modeling

*Chemometric Models*

Chemometric Models that were considered were partial least squares (PLS) regression, support vector regression (SVR), convolutional neural networks (CNN), and Light Gradient Boosting Model (LightGBM). We would explore using chemometric models to find relationships between the spectral data and the soil properties with the calibration data in mind.

*Partial Regression*

Initially, PLS was run and proved to be one of the best-performing, simplest models. This model is advantageous as it has built-in dimensionality reduction and uses latent variables to make predictions. Four principal components were selected from tuning the hyperparameters with an RMSE = 0.2895. Due to limitations with available hyperparameter tuning options with PLS, it was determined to seek out other models to achieve better accuracy that include more tunable parameters.

*Support Vector Regression*

SVR was explored as another model. This model attempts to find a hyperplane in the dataset that results in a minimum error between training and predicted values, according to the loss function, RMSE. SVR can handle complex datasets and account for linear, polynomial, and non-linear (RBF) relationships. As opposed to the previous model, quite an array of hyperparameters can be optimized for this algorithm. The best accuracy achieved with SVR was RMSE = 0.2916. As promising as this model was, it was eventually withdrawn from consideration due to hyperparameter optimization expending substantial amounts of time. In one instance, two days were spent attempting to process optimization code in Python and it was eventually terminated.

*Convolutional Neural Networks*

Convolutional Neural Networks is a deep learning framework often used in image classification and works well for hyperspectral matrices. Another benefit of CNN is the potential for customization of hyperparameters and layers that can be ideally configured to the dataset. This model achieved better results, with an RMSE = 0.2885, and runtimes were quite manageable. Extensive experimentation consisted of adding and removing batching, pooling, and fully connected layers. In addition, hyperparameters such as kernel size, number of units, number of filters, learning rate, and epochs were configured for the model using the RandomSearchCV algorithm. This algorithm takes a specified random sample of the pool of hyperparameters and iteratively determines what settings score best against a specified loss function.

*Light Gradient Boosting*

The last chemometric model to be evaluated was the LightGBM regression model. This model required four separate models for each soil parameter prediction, as multivariate models were impossible. It allowed for more customizable parameter tuning per soil parameter. The Random Search algorithm was employed for 11 hyperparameters and achieved an RMSE = 0.2884, the best-performing model.

Distributions of training soil parameter values compared to predicted values were plotted, as seen in Figures 2-4. Except for phosphorus, each soil parameter's distribution shape and average mean exactly mimics training data. Phosphorus still bears a strong resemblance to training's

distribution, although its right skew is much stronger. The difference in thickness is due to less test data than training data.

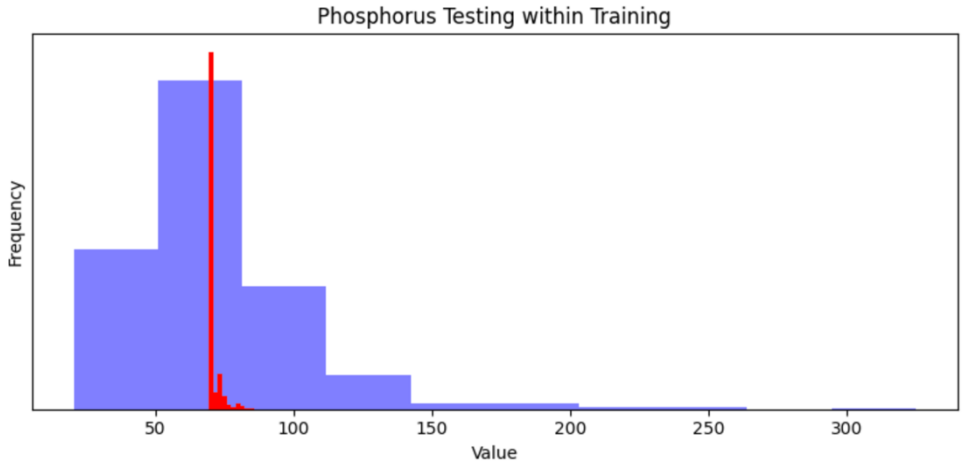


*Figure 6*

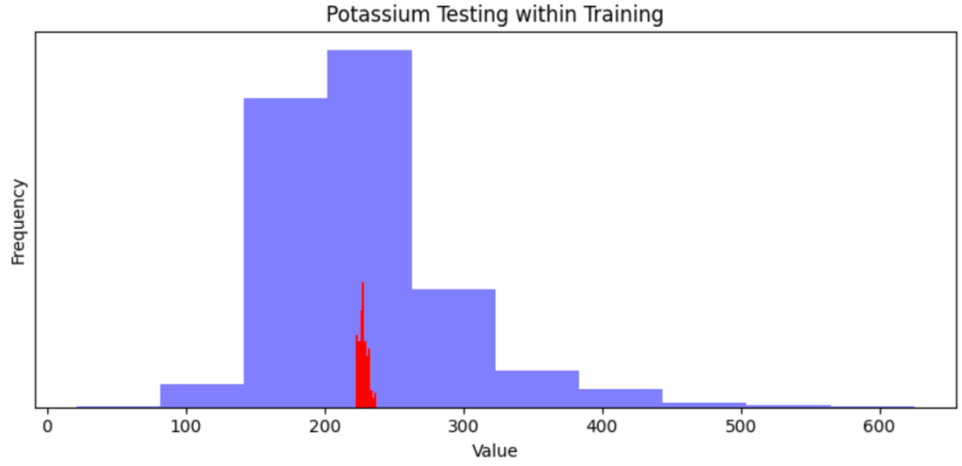


*Figure 7*

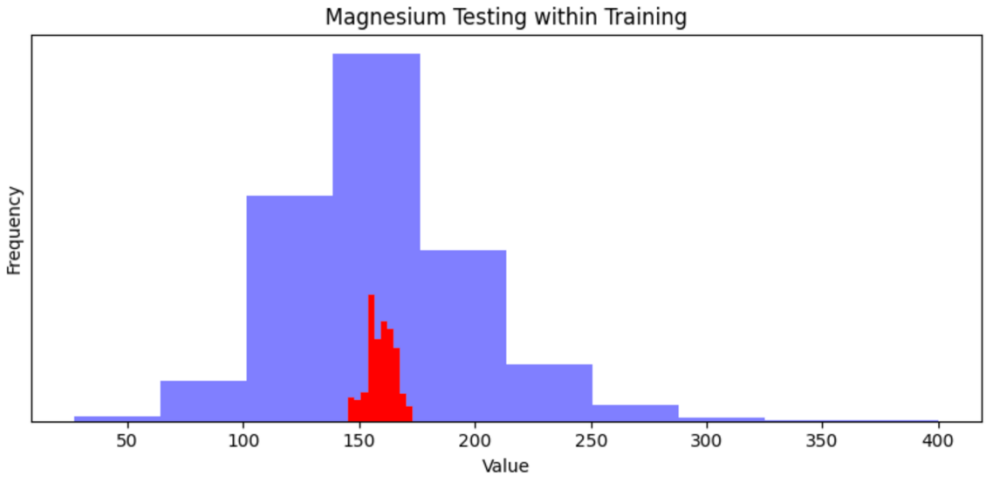


*Figure 8*

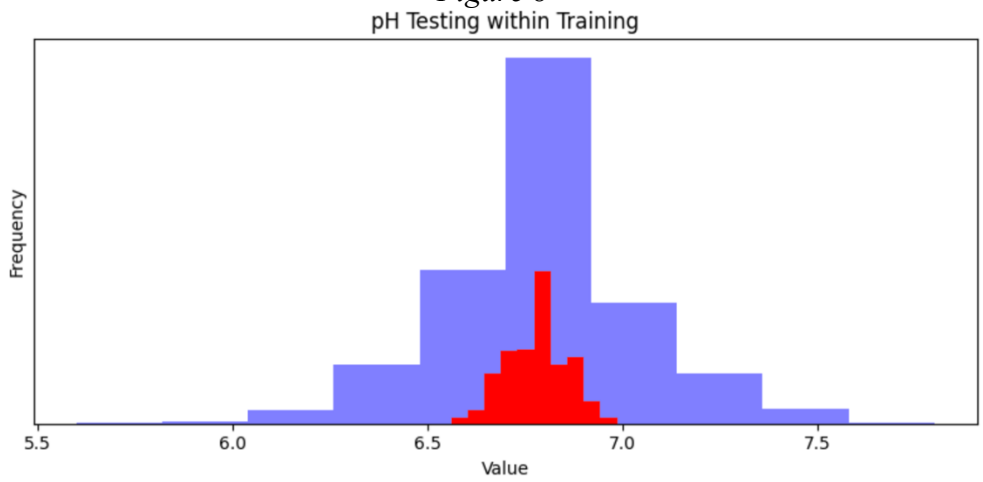


*Figure 9*

The fact that the test data reflects the training data can be two-fold: one aspect is that the model overfits and mimics training data too closely. The second is that it is reflective of real-world, empirical values. To further generalize the model, additional training data could help to lower errors in predictions associated with overfitting.

Several strategies were employed to encourage model generalization, such as limiting depth, L1/L2 regularization, and minimum child samples. The following table shows hyperparameters for the LightGBM soil parameter models.

| Hyperparameter | P | K | Mg | pH |
|---|---|---|---|---|
| Number Leaves | 3 | 12 | 12 | 12 |
| Number Estimators | 40 | 40 | 20 | 40 |
| Max Depth | 5 | 9 | 5 | 5 |
| Learning Rate | 0.03 | 0.01 | 0.05 | 0.05 |
| Boosting Type | dart | dart | gbdt | gbdt |
| Bagging Frequency | 5 | 7 | 3 | 1 |
| Bagging Fraction | 0.75 | 0.9 | 0.75 | 0.9 |
| L1 Regularization | 0.3 | 0.1 | 0.5 | 0.3 |
| L2 Regularization | 0.2 | 0.1 | 0.3 | 0.2 |
| Min Child Samples | 20 | 100 | 100 | 20 |
| Feature Fraction | 0.9 | 0.6 | 0.75 | 0.75 |

*Table 1*

# Feature Importance

To assess feature performance, SHAP (SHapley Additive exPlanations) was employed to understand better what features most impact each soil parameter and to what degree. SHAP allows one to assess a specific row instance and see each feature's contribution to the model's prediction. In this way, the granular decision framework within the model is revealed. Three plots were made to visually characterize the relationships the LightGBM model exploits to make predictions. A bar chart with feature importance is plotted against 150 spectral bandwidths from the dataset for each parameter. It gives a quick picture of what features stand out and where on the spectrum they are positioned.

Next, a summary of the plot lists features in descending order of importance. It is important to note that some of the tallest peaks in the former plot may be in a lower-than-expected rank on the summary plot. It may be due to interactive effects between predictors that lessen or heighten their importance in the model.

The final plot is forced. It helps understand the individual marginal contribution of each feature in the prediction value for a given data point as the closer to the centerline, the more critical, with the respective bar's width proportional to its effect's magnitude.

*Phosphorus*

The bar plot below shows feature importance peaks for phosphorous and their associated color if on the visible spectrum. There were notable spikes on the far left with tight groupings in orange and red.

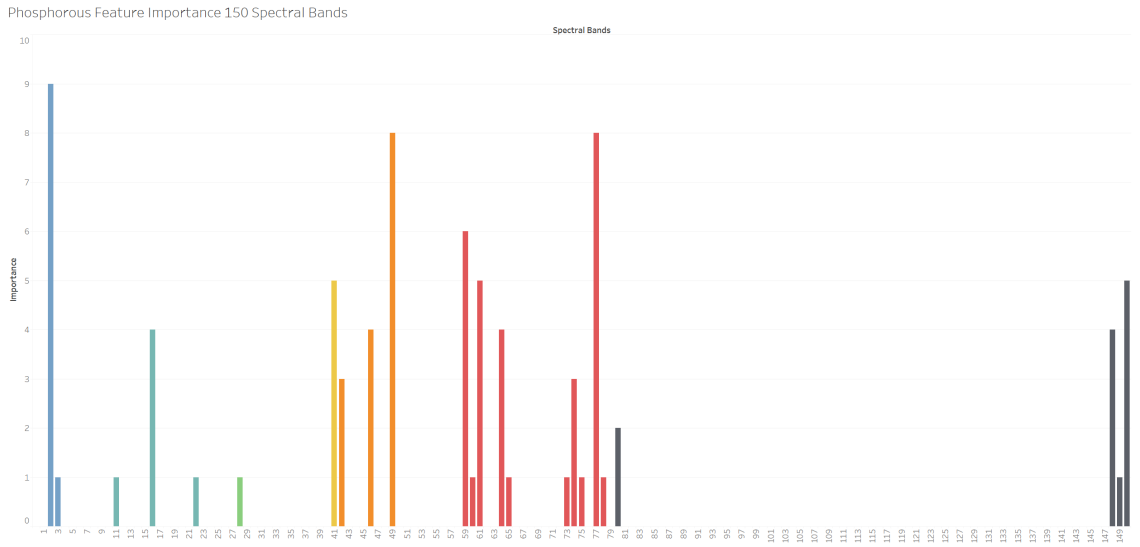


*Figure 10*

The top features are ranked below in the summary plot. Notably lower in this plot is the far-left spike. It may be due to an interactive effect; its importance is lessened when combined with another feature.

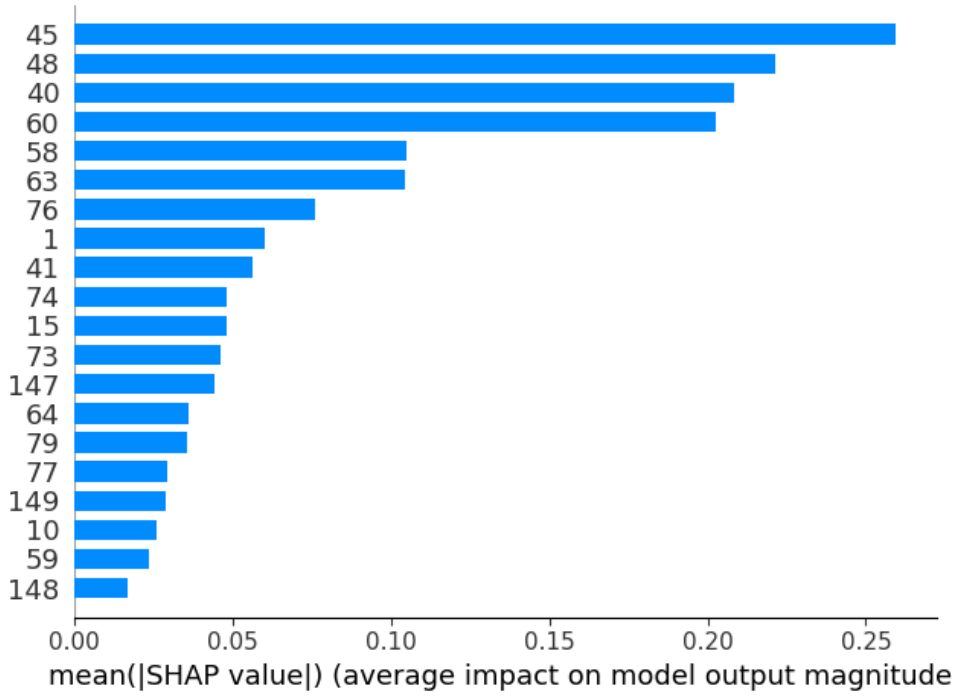


*Figure 11*

The force plot for phosphorous for a specified data point reveals a predicted value of 69.53 ppm. The baseline value is about 70.25 ppm, and some features, colored blue, lower this value to its expected value, as the red values contribute towards a higher predicted value. It is like a regression equation with negative or positively associated parameters totally to a final value.

It is worth noting that Feature 44 in the force plot correlates to Feature 45 in the summary plot and so on due to how the Python library handles the indexing. In this case, the top feature in the previous plot displays the most significant magnitude towards increases in the predicted phosphorous concentration. The following features concertedly work to reduce the concentration, namely Features 47 and 39.

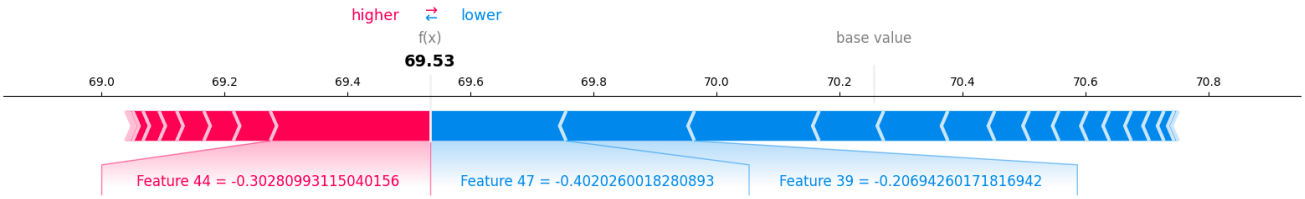


*Figure 12*

*Potassium*

Potassium has overtly complex relationships with predicting its concentration. It can be seen from the number of variables displayed in the bar chart below used to indicate this target. There are noticeable peaks near bandwidth numbers 1, 35, and 69.

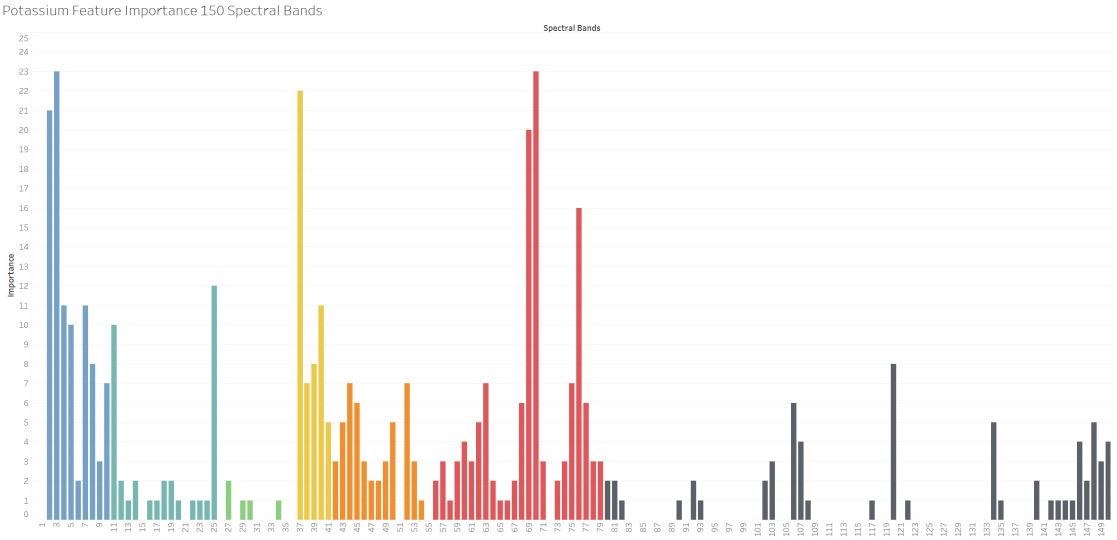


*Figure 13*

The summary plot validates the visual findings that 69 is the most crucial feature, with 68 just below. There is a noticeable drop in importance magnitude after these bandwidths. The high incidence of bandwidth response and drop in importance metric indicates the likelihood of interactive effects affecting the predictive response of the target variable.

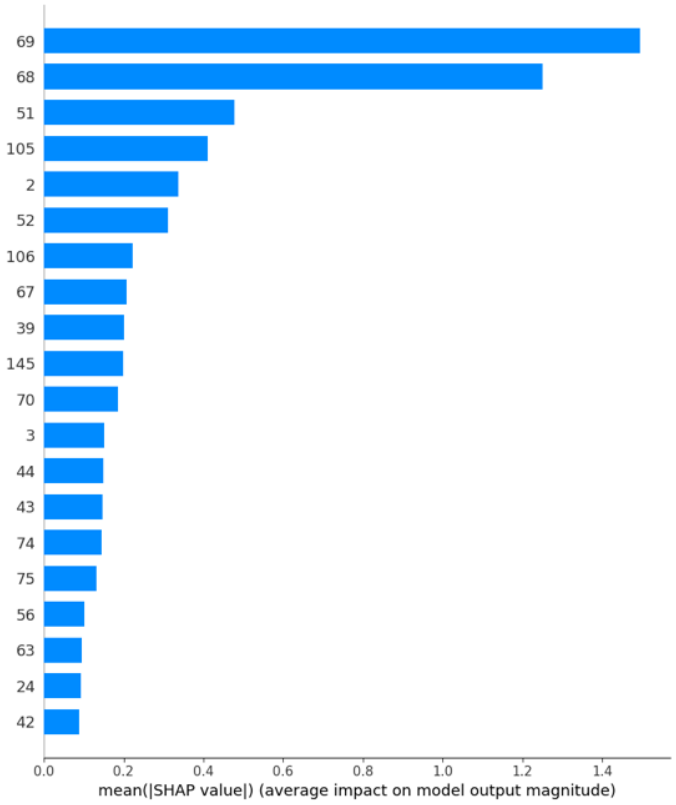


*Figure 14*

The force plot reveals Features 69 and 68 have the most significant magnitude in altering the predicted amount of potassium concentration. These bandwidths serve to lower the baseline value. The remnant of features, though individually small, adds significant contributions to raise and lower the target amount.

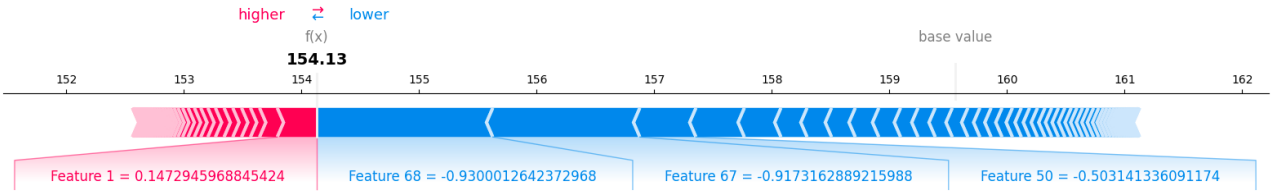


*Figure 15*

*Magnesium*

A visual inspection of Magnesium in the bar chart below reveals similar characteristics of Potassium, with many features at play in predicting. There are peaks near bandwidth 3, 68, and 150.

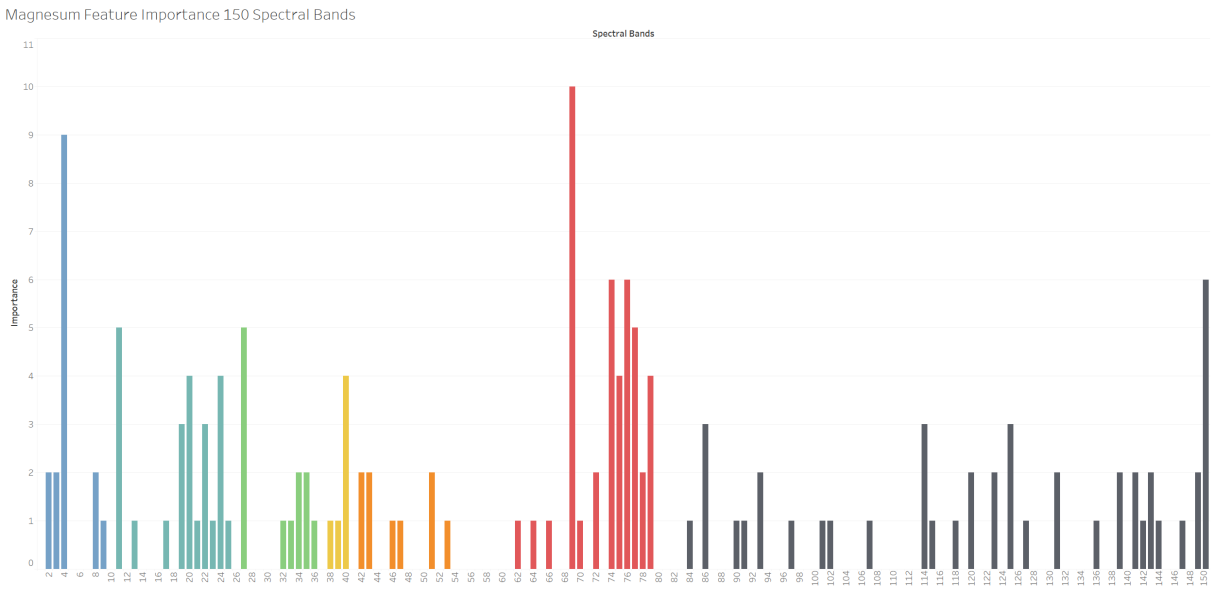


*Figure 16*

The summary plot reveals existing dynamics that interact to predict the magnesium concentrations in soil. It is observed because none of the peak values identified in the bar chart stand out in the summary plot. The top features in the summary plot have a minor visual representation in the bar chart, pointing to interactive solid effects.

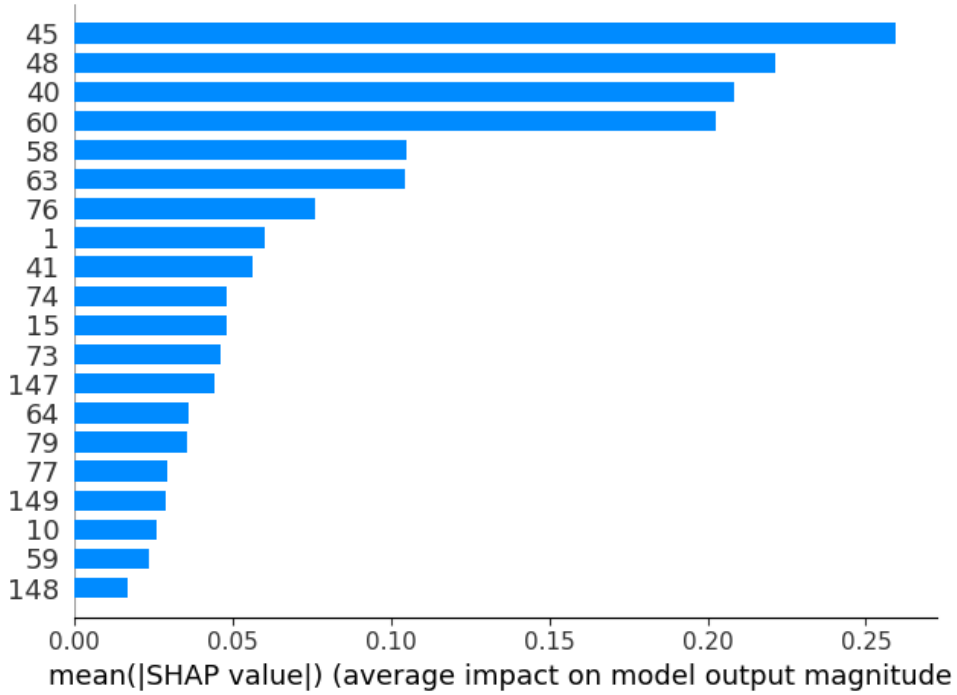


*Figure 17*

The force plot reveals a symmetrical framework where both positive and negative effects counteract each other evenly. This pattern may show the actual decision network for this class, or it could be due to low responsiveness to the model in predicting this target variable; however, considering the diverse array of bandwidths in the bar chart, a complex balance of features influences this parameter. So far, other parameter predictions have appeared to have more specialized features, yet Magnesium requires a broad bandwidth reading for accurate quantification.

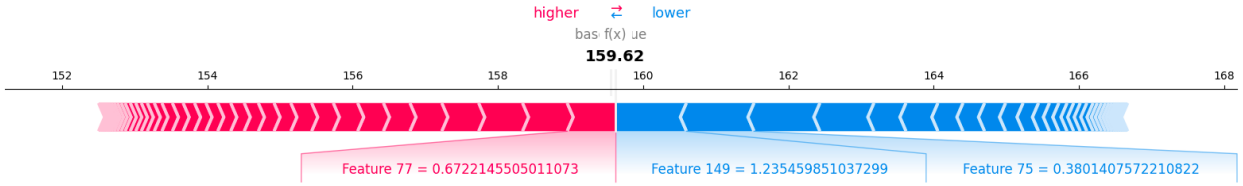


*Figure 18*

*Soil pH*

The last soil parameter that was characterized was pH. The leftmost bandwidths stand out as the highest magnitude with a significant grouping around Feature 70.

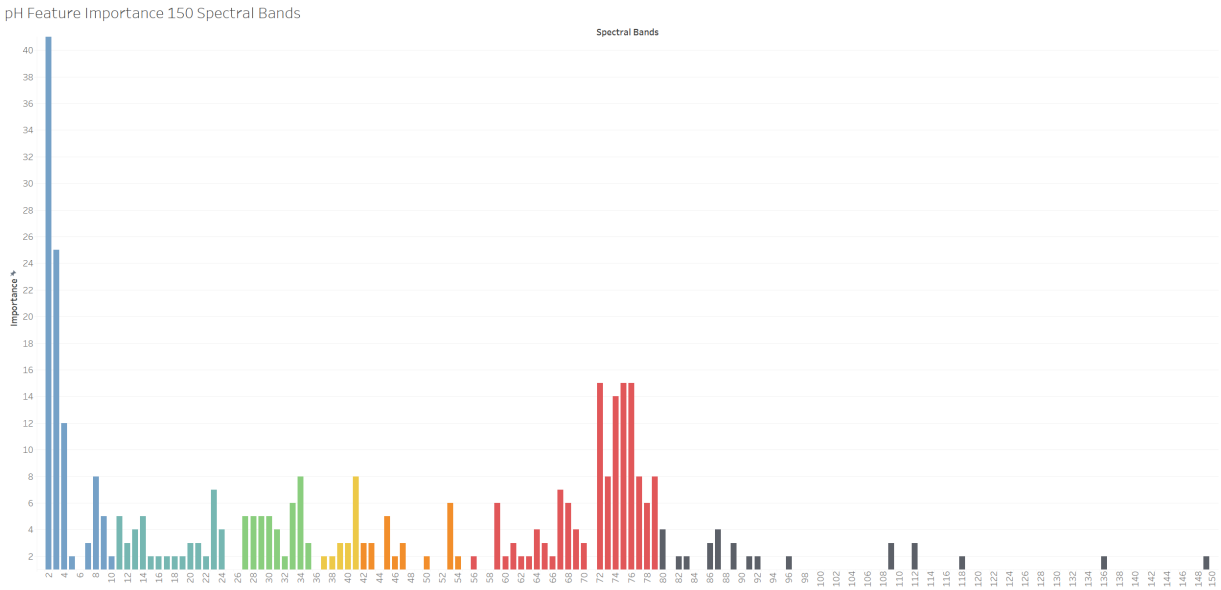


*Figure 19*

From the summary plot, feature 73 appears to be the most significant and is positioned near the second-highest cluster in the plot above. It indicates that its interactive relationship(s) exceeds any other feature's contribution. The tall spike on the far left is also included towards the top of this list, Feature 1.

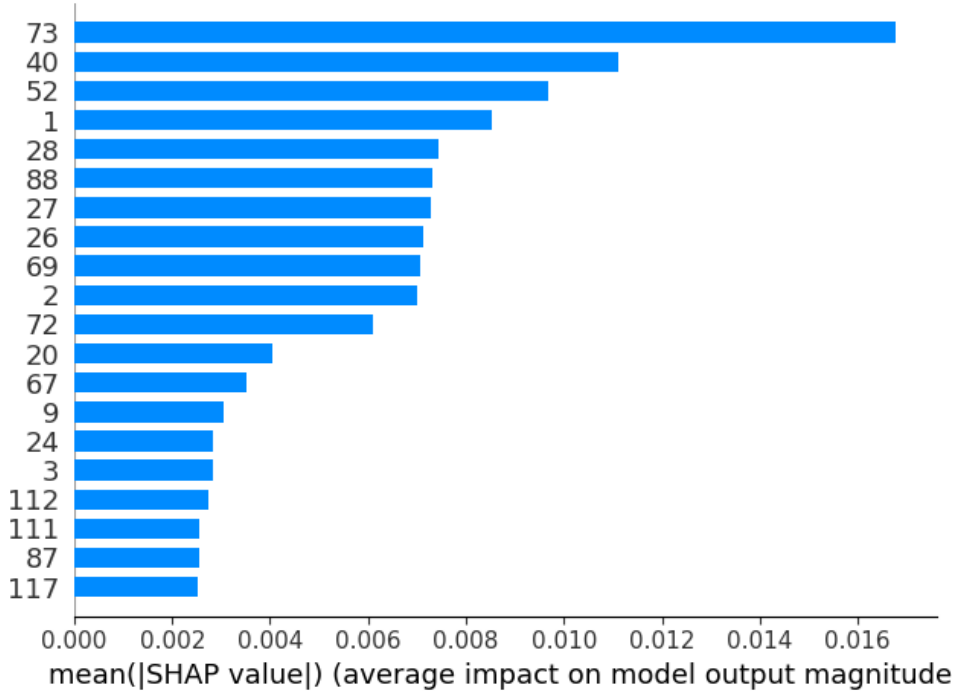


*Figure 20*

The force plot further confirms the findings discussed in the summary plot. Feature 71 has the most significant contribution, stemming from the interactive effect, as its effect appears insignificant. Aside from this, many features contribute to raising the value of the pH term, making it more basic.

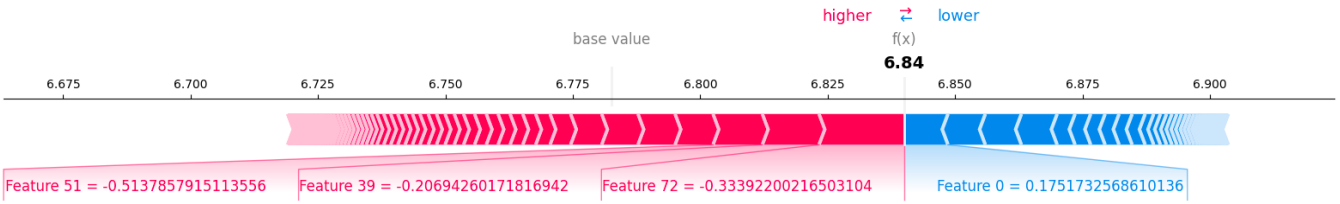


*Figure 21*

# Insights & Recommendations

*Insights & Recommendations for Phosphorous*

The ability to administer phosphorous with clinical precision is vital to crops and water bodies. Not only this, according to a National Geographic article, "All of the phosphorus that farmers use today—and that we consume in the food we eat—is mined from a few sources of phosphate rock, in the United States, China, and Morocco. By some estimates, those could run out in as little as 50 to 100 years." Excess phosphorous causes algal blooms that can cause devastating consequences to aquatic life. Therefore, research that attempts to measure regional surface water health changes after adopting hyperspectral model predictions could contribute to innovation in this field.

Looking at the SHAP force plot for phosphorous, some bandwidths were inversely related and close to one another. Bandwidths 44 and 47 were the most critical features that impacted the phosphorous calculation. It is understood that measurement accuracy, calibration, and resolution are of chief importance. The ability to specialize phosphorous reading between Bandwidth 39 to 48 with higher resolution would be advantageous to tuning predictions for this soil parameter.

*Insights & Recommendations Potassium*

According to the Climate Change Post, adequate potassium in soil can improve water use efficiency in plants, which is particularly important in regions with water scarcity. Poland faces water scarcity and has one of Europe's lowest freshwaters (lakes and rivers) availabilities. As Poland is under medium to high stress, this research is pivotal in water conservation efforts, especially the analysis of potassium. Further research is recommended to assess whether widespread hyperspectral-based modeling recommendations help improve water efficiency in agricultural practices.

Regarding potassium's SHAP summary and force plots, the spectral bandwidth ranges from 50 to 70, which is of primary importance to this soil parameter. As with the previous recommendation, further enhancements to precision in this range enable better predictions and the ability to characterize hyperspectral relationships better.

*Insights & Recommendations for Magnesium*

According to a study by Qadir et al., water scarcity necessitates the effective use of soil. "High-magnesium waters and soils are emerging examples of water quality deterioration...soil degradation and impact crop yield negatively." In the LightGBM model, there was shared importance of many features across a broad spectrum ranging from both ends. It supports the need for wide-spectrum measurements.

*Insights & Recommendations for pH*

From the model, many features contribute to making pH for alkaline except Bandwidth 0. Most are in the 39 to 72 bandwidth range or yellow to red visible spectrums. Enhancements in this range will further contribute to better predictions for soil pH.

For all soil parameters besides magnesium, the bandwidth ranges from 39 to 72 is highly significant in making predictions. In addition, visible light spectral readings carry more importance than the higher, not-visible bandwidths, though these still weight some parameters. Magnesium possessed the most unique characteristics in terms of attempting to characterize the relationship between hyperspectral readings and this parameter. Further model refinements or deep learning frameworks trying to describe this complex interplay better may reveal more insights.

Hyperspectral technology paired with powerful machine learning techniques is pivotal in tackling modern challenges with water scarcity, food insecurity, and environmental impacts. Enhancing resolution and ensuring accuracy within the bandwidth range of 39 to 72, or 583.55 to 689.03 nanometers, can significantly refine the specialized characterization of soil parameters, thereby improving prediction accuracy.

Looking at the overall landscape of hyperspectral imaging, being able to scale and market hyperspectral imaging has been a common issue in the past. For the system developers and manufacturers, taking the models, charts, and findings that we were able to garner is only half of the battle, as being able to translate it to everyone universally is the other half of the battle. Not

being able to market and scale the hyperspectral imaging along with the other findings will make all the work go to waste, so being able to translate the work that has been done correctly is the priority so that hyperspectral imaging can be at the forefront so that everyone can understand and utilize it.

# References


"Fresh Water Resources in Poland." *Climatechangepost.Com*, www.climatechangepost.com/poland/fresh-water-resources/. Accessed 6 Nov. 2023.

"Global Hyperspectral Imaging Market Report 2022-2030: Rising Need for in-Depth Data from the Optical Images Will Increase Demand for Hyperspectral Imaging Systems." *GlobeNewswire News Room*, Research and Markets, 1 Apr. 2022, www.globenewswire.com/en/news-release/2022/04/01/2414594/28124/en/Global-Hyperspectral-Imaging-Market-Report-2022-2030-Rising-Need-for-In-depth-Data-from-the-Optical-Images-will-Increase-Demand-for-Hyperspectral-Imaging-Systems.html. Accessed 10 Nov. 2023

Qadir, M;Schubert, S;Oster, JD;Sposito, G;Minhas, PS;Cheraghi, SAM;Murtaza, G;Mirzabaev, A;Saqib, M; (n.d.). *High-magnesium waters and soils: Emerging environmental and food security constraints*. The Science of the total environment. https://pubmed.ncbi.nlm.nih.gov/30045492/#:~:text=A%20ratio%20of%20magnesium%2Dto,and%20impact%20crop%20yields%20negatively. Accessed 10 Nov. 2023.

Rosen, Julia. "Farmers Are Facing a Phosphorus Crisis. the Solution Starts with Soil." *Science*, National Geographic, 3 May 2021, www.nationalgeographic.com/science/article/farmers-are-facing-a-phosphorus-crisis-the-solution-starts-with-soil. Accessed 10 Nov. 2023.

Specim. "What Is Hyperspectral Imaging?: A Comprehensive Guide - SPECIM Spectral Imaging." *Specim*, 8 Aug. 2023, www.specim.com/technology/what-is-hyperspectral-imaging/. Accessed 6 Nov. 2023.