Investigating the Use of Graph Convolutional Networks on Soil Phosphorus Estimation from Hyperspectral Sensor Data

N. Pelin Metin[a], Yucel Cimtay[b], Nazila Panahi[c]

[a]Politecnico di Milano, Computer Science and Engineering, Milan, Italy

[b]TOBB ETU, Computer Engineering, Ankara, Turkey

[c]Urmia University, Electrical Engineering, Urmia, Iran

Abstract. Accurate and rapid prediction of soil phosphorus, a vital nutrient for agricultural productivity, is crucial for optimizing fertilization practices and reducing environmental impacts. While traditional laboratory analyses are reliable, they are often time consuming and costly. In this study, a graph-based deep learning approach is proposed to estimate soil phosphorus content using hyperspectral data acquired from ground based sensors. Hyperspectral measurements from the *Spectroscopic data of soil samples doped with* P_2O_5 dataset were preprocessed through data augmentation and Min–Max scaling, followed by dimensionality reduction using Mutual Information Regression to retain the most informative spectral bands. Soil samples were modeled as nodes in a graph constructed via the k-nearest neighbors algorithm to capture spectral relationships, and a Graph Convolutional Network consisting of two convolutional layers and a fully connected output layer was trained using 3-fold cross validation. The latest cosine similarity search confirms the original 300 band manuscript configuration as the main setting, reaching mean validation $R^2 = 0.8779$ and mean test $R^2 = 0.8465$. These results demonstrate that integrating hyperspectral sensing with graph neural networks effectively captures complex spectral dependencies and provides a promising tool for precise and sustainable soil nutrient management.

Keywords: Graph Neural Network (GNN), Soil Phosphorus Estimation, Hyperspectral Sensor Data, Graph Convolutional Networks (GCN), Machine Learning (ML).

1 Introduction

Hyperspectral imaging (HSI) has emerged as a powerful sensing technology that captures detailed spectral information across hundreds of contiguous wavelength bands, enabling the identification of material specific spectral signatures. In recent years, the integration of HSI with artificial intelligence (AI) and machine learning (ML) techniques has significantly improved the ability to monitor soil properties, vegetation health, nutrient status, crop stress, and other agricultural variables.

Recent advancements in molecular techniques have greatly enhanced the understanding of phosphorus (P) metabolism and transport mechanisms in plants. This essential macronutrient is vital for regulating plant growth, development, and reproduction. Phosphorus is closely related to root formation, flowering, seed production, and energy transfer, so fast and reliable phosphorus monitoring is important for sustainable fertilizer management.

Traditional methods for assessing phosphorus include various laboratory chemical tests, sensors, and infrared spectroradiometers. Although these methods are accurate and well established, they are expensive, labor intensive, and difficult to repeat at high spatial or temporal resolution. These drawbacks have led researchers to implement data-driven models for predicting soil characteristics from spectral measurements.

Estimating soil phosphorus from hyperspectral data is a challenging task due to several inherent factors. First, hyperspectral measurements are characterized by extremely high dimensionality, with strong inter-band correlations and redundant information. Second, phosphorus does not appear as a single isolated spectral absorption feature; instead, it is indirectly expressed through interactions with other soil properties. Therefore, models must extract subtle patterns from many correlated bands while avoiding overfitting on small datasets.

Traditional machine learning and deep learning approaches, such as CNNs, primarily operate on individual samples or local spectral patterns, without explicitly modeling the relationships between different samples in the dataset. To address these challenges, this study adopts a Graph Convolutional Network (GCN) framework, where each soil sample is represented as a node and edges encode similarity relationships in the spectral feature space.

2 Related Work

Hyperspectral imaging (HSI) captures reflected or emitted radiation across many narrow, continuous bands. Unlike RGB imaging, which records only three broad bands, hyperspectral sensing provides a detailed spectral signature for each sample. Recent reviews show that HSI has become a major tool in precision agriculture, supporting crop monitoring, disease detection, weed identification, soil mineral analysis and nutrient assessment [7, 3]. For soil studies, this dense spectral information is valuable because reflectance patterns can encode indirect information about organic matter, texture, moisture and nutrient status. This is particularly important for phosphorus estimation, where the spectral response is often indirect rather than tied to a single dominant wavelength.

Recent spectroscopy studies demonstrate the potential of spectral sensing for phosphorus and broader nutrient estimation. Wali et al. used photoacoustic Fourier transform mid-infrared spectroscopy to estimate plant available

phosphorus in bio-based fertilizers and amended soils [11]. Wan et al. predicted soil organic matter, available nitrogen, available phosphorus and available potassium using a nearby hyperspectral sensor under varying soil moisture conditions [12]. Rivadeneira-Bolaños et al. provided the phosphorus reflectance dataset used in this work, showing that machine learning models can be applied to laboratory spectra for phosphorus prediction [8]. Nalepa et al. introduced the HYPERVIEW benchmark, which combines airborne hyperspectral imagery with in situ measurements of phosphorus pentoxide, potassium oxide, magnesium and pH [6]. Together, these studies indicate that spectral sensors can support nutrient prediction, but performance depends strongly on calibration, soil conditions, sample size and preprocessing choices.

Machine learning and deep learning methods are increasingly used to improve soil NPK sensing and nutrient prediction [1]. Deep models can learn nonlinear spectral patterns from high-dimensional inputs, and recent reviews highlight CNN, RNN, LSTM and transformer architectures for hyperspectral agriculture problems [3]. These approaches are useful for extracting spectral features, but they often require careful dimensionality reduction because hyperspectral bands are highly redundant and correlated. Band selection is therefore an important preprocessing step, especially in small datasets where using all bands can increase overfitting risk. It also helps focus the model on wavelength regions with stronger predictive relevance.

Despite these advances, many hyperspectral pipelines process each spectrum as an independent sample or as a one dimensional sequence. This can limit their ability to use relationships between spectrally similar soil samples. Graph Neural Networks (GNNs) provide a natural framework for modeling such relationships. The original GNN formulation introduced neural computation over graph structured data [10], while GCNs made neighborhood aggregation practical for learning node representations [5]. In hyperspectral analysis, graph based methods have been used to combine spectral and spatial context for classification and representation learning [13, 2]. Hong et al. [4] also demonstrated the effectiveness of Graph Convolutional Networks for hyperspectral image classification, where graph connections can represent relationships among spectrally similar pixels. This supports the general relevance of graph-based learning for hyperspectral data analysis. However, unlike image classification studies, the present work focuses on regression from isolated soil spectra rather than pixel-level hyperspectral images.

The present study builds on this direction by constructing a graph over soil samples rather than image pixels. Since phosphorus is not expressed by a single isolated spectral feature, a graph over spectrally similar samples may provide useful contextual information beyond each individual spectrum. This motivates the use of GCNs for phosphorus estimation from hyperspectral sensor data. Therefore, the graph in this study should be interpreted as a spectral similarity graph between independent soil samples. Since the available dataset does not contain hyperspectral images or spatial coordinates, spatial neighborhood information is not included in the current model.

3 Materials and methods

3.1 Dataset

This study utilizes the dataset described in [8], which comprises 152 reflectance spectra ranging from 200 to 900 nm. These spectra represent soil samples prepared with varying concentrations of P_2O_5, collected from Santander de Quilichao, Cauca, Colombia. The local project files contain `data.xlsx` with 3501 spectral bands by 152 samples and `labels.xlsx` with 152 phosphorus measurements.

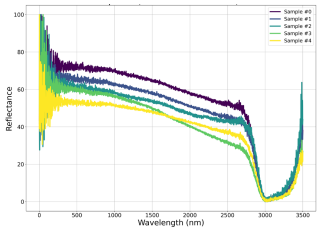


Figure 1: Example reflectance spectra for soil samples with different phosphorus levels.

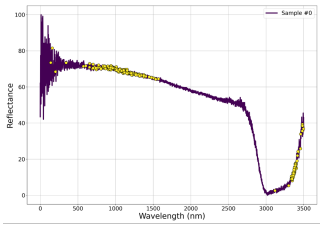


Figure 2: Representative spectrum with mutual information selected bands highlighted.

The phosphorus measurements cover a very broad range, spanning almost three orders of magnitude. In practice, this means that the dataset includes soils from very low to very high phosphorus levels. The broad range is useful for regression because it includes diverse phosphorus conditions, but it also makes prediction difficult because errors at high values can dominate the loss and the model must generalize from only 152 samples.

3.2 Method

Graph Convolutional Networks were employed to estimate soil phosphorus concentration from hyperspectral data while using graph edges to represent relationships between spectrally similar samples. In message passing, each node aggregates information from its neighbors and updates its embedding according to both its own features and the graph structure.

In this paper, each soil sample is represented as a node within a graph structure, while similarities between samples in the spectral feature space are encoded as edges. The graphs are generated using a k-nearest neighbor procedure based on spectral similarity. The value of k controls the level of connectivity and the extent of information propagation within the graph: very small values may produce a poorly connected graph, whereas very large values may lead to oversmoothing by linking samples with substantially different spectral characteristics.

3.3 Preprocessing

To improve generalization, Gaussian noise was added to the training set, generating six synthetic samples from each training sample. The noise was scaled by the distribution of each spectral band, and negative reflectance values were clipped.

In this work, additive Gaussian noise was the only data augmentation strategy applied. This choice was made to simulate small sensor-level perturbations while preserving the overall spectral shape of each sample. Other augmentation strategies could be considered in future work, such as multiplicative intensity scaling, baseline shifts, wavelength jitter, random band masking, spectral smoothing perturbations, or mixup between chemically similar samples.

The spectral data were normalized with Min-Max scaling to map each spectrum to a fixed range while preserving spectral shape and reducing scale differences between samples. An $\varepsilon = 10^{-8}$ term was used to avoid division by zero.

Because hyperspectral data contain thousands of redundant and highly correlated bands, Mutual Information Regression [9] was used to select informative bands and reduce overfitting. The initial manuscript setting used 300 selected bands; later band count experiments over 200, 250 and 300 bands confirmed 300 as the main configuration. This criterion is suitable because it can capture nonlinear dependencies between each band and phosphorus concentration.

3.4 Proposed Architecture and Evaluation

The proposed GCN architecture operates on a graph constructed using the k-nearest neighbors algorithm, where each node represents a soil sample and edges reflect spectral similarity between samples. The input features consist of spectral bands selected from the 3501 original bands using Mutual Information Regression after Min-Max Scaling. During training and validation, the model takes the augmented, preprocessed and band selected reflectance vector of each soil sample as input, where the output is the estimated amount of phosphorus of the corresponding soil sample.

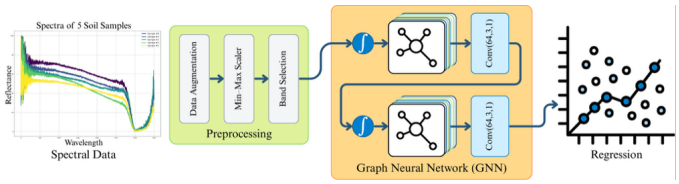


Figure 3: Proposed GCN pipeline for soil phosphorus estimation, including preprocessing, graph convolution and regression.

The GCN model is implemented with two graph convolution layers, each with e hidden units, followed by a fully connected output layer. Each convolutional layer is followed by a ReLU activation and a dropout layer with a dropout rate of 0.3 to prevent overfitting. The final linear layer maps the e-dimensional node embeddings to a single continuous output corresponding to predicted phosphorus concentration. Model parameters are optimized using Adam with weight decay 10^{-4} and mean squared error loss.

Table 1: Hyperparameter settings used in GCN experiments.

Hyperparameter	Candidate values / final focus
Selected bands	200, 250, 300; manuscript setting: 300
Embedding size	64, 128, 256; 300-band setting: 128
Maximum epochs	500, 1000, 1500; 300-band setting: 1000
Neighbors k	3, 5, 7; latest best: 3
Augmentations	5, 6, 7; 300-band setting: 6
Learning rate	0.001
Noise scale	0.075, 0.1; 300-band setting: 0.075

Model performance was assessed and reported using the coefficient of determination (R^2) on validation and test folds.

4 Results

After applying the proposed GCN architecture to the preprocessed reflectance data, an extensive hyperparameter search was applied. For each hyperparameter combination, the model was trained and evaluated by 3-fold cross validation, and the coefficient of determination was computed on the held-out folds. The previous Euclidean GCN result is kept because it explains what changed from the older manuscript, while the updated cosine similarity experiments confirm the 300 band manuscript setup as the main configuration.

Table 2: Performance summary for phosphorus estimation. Values are mean R^2 scores unless otherwise noted.

Model / setting	Bands	Val. R^2	Test R^2
GCN, Euclidean k-NN	300	0.8195	0.7307
GCN, cosine k-NN, manuscript setting	300	0.8779	0.8465
GCN, cosine k-NN, 200-band comparison	200	0.8623	0.8496
1D-CNN benchmark	300	0.75	0.76
LSTM benchmark	–	0.89	0.89
ANN benchmark	–	0.88	0.84
Random Forest benchmark	–	0.97	0.88

Among the tested GCN configurations, the updated 300 band cosine similarity setting uses embedding size 128, maximum epochs 1000, patience 40, $k = 3$, six augmentations, learning rate 0.001 and noise scale 0.075. On the validation folds this model reached mean $R^2 = 0.8779$, and on the independent test folds it achieved mean $R^2 = 0.8465$. The 200 band comparison was a near tie on test R^2 but lower on validation, so the experiments confirmed the original 300 band setting as the main manuscript configuration.

Table 3: Fold-level results for the latest 300-band cosine-similarity GCN configuration.

Fold	Validation R^2	Test R^2
1	0.8622	0.8940
2	0.9223	0.7952
3	0.8493	0.8504
Mean	0.8779	0.8465

The remaining top GCN configurations obtained slightly lower performance, indicating that training length, patience, neighborhood size and augmentation noise affect robustness on this small dataset. Against the project baselines, Random Forest shows a high validation score but a visible validation-test gap, while the GCN is better than CNN and competitive with LSTM and ANN.

5 Ablation Study and Discussion

Without applying band selection, using all bands made the estimation results worse. The mean R^2 score reached only about 0.51, showing that band selection is a required and meaningful step for handling more accurate results. This confirms that using all 3501 bands increases redundancy and overfitting risk in such a small dataset.

The KNN part of the GCN model is very important in terms of determining the correct points used as the nodes. When the metric was changed from the default Euclidean distance to cosine similarity, the results improved. Cosine similarity better fits this study because it compares the direction, or spectral shape, of vectors rather than their magnitude. This makes it robust to variations in illumination, sensor gain, or overall brightness, which are common issues in hyperspectral measurements.

In an operational setting, the choice between isolated spectra and hyperspectral imaging depends on the target application. If the objective is rapid laboratory or point-based soil screening, isolated spectra can be sufficient. However, if the goal is to map phosphorus variability across a soil surface or agricultural field, hyperspectral imaging would be more appropriate. In that case, spatial information could be incorporated by defining graph nodes as pixels, superpixels, or sampling locations, and by connecting them according to both spectral similarity and spatial proximity.

Prior spectroscopy studies report mixed phosphorus prediction accuracy across soil types and preprocessing strategies. The present results support the idea that graph construction and band selection can improve prediction when phosphorus is indirectly represented in the spectrum.

The robustness of the findings was assessed through 3-fold cross-validation, comparison with baseline models, ablation of band selection, and sensitivity analysis over graph construction choices such as the distance metric, number of neighbors, number of selected bands, and augmentation noise level. Nevertheless, stronger robustness assessment would require repeated cross-validation, confidence intervals or bootstrap analysis, external validation on independent soils and sensors, and perturbation tests under simulated noise, wavelength shifts, moisture variation, and calibration differences.

The main limitation is dataset size. With only 152 samples, fold-level variation is expected. The best cosine model has strong average performance, but one test fold is lower than the other two. Larger datasets, additional soil types and independent field measurements are needed before claiming broad operational readiness. Still, the present results show that GCNs are a promising modeling direction for hyperspectral nutrient estimation.

6 Conclusions

This study demonstrated the applicability of Graph Convolutional Networks for estimating soil phosphorus from hyperspectral measurements. The final pipeline combines data augmentation, Min-Max scaling, Mutual Information Regression band selection, k-nearest-neighbor graph construction and two graph convolutional layers trained with 3-fold cross-validation.

The obtained performance demonstrates that the GCN-based approach can capture spectral relationships relevant to soil phosphorus. The model is not intended to replace laboratory analysis in all situations, but it can support faster screening and decision-support systems for precision agriculture when integrated with field or near field sensors. Future work should test larger soil datasets, additional nutrients, and alternative graph definitions such as weighted graphs, adaptive neighborhoods and graph attention mechanisms.

Disclosures

NPM contributed data curation, visualization, code development and editing. NP contributed related work writing and literature review. YC contributed supervision, data curation and original drafting. The authors declare no funding and no known competing interests.

References

[1] S. Dattatreya, A. N. Khan, K. Jena, and G. Chatterjee. Conventional to modern methods of soil npk sensing: A review. *IEEE Sensors Journal*, 24(3):2367–2380, 2024.

[2] S. Geisler, A. Kosmala, D. Herbst, and S. Günnemann. Spatio-spectral graph neural networks. In *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS 2024)*, Vancouver, 2024.

[3] M. F. Guerri, C. Distante, P. Spagnolo, F. Bougourzi, and A. Taleb-Ahmed. Deep learning techniques for hyperspectral image analysis in agriculture: A review.

ISPRS Open Journal of Photogrammetry and Remote Sensing, 12:100062, 2024.

[4] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot. Graph convolutional networks for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 59(7):5966–5978, 2021.

[5] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[6] J. Nalepa et al. Estimating soil parameters from hyperspectral images: A benchmark dataset and the outcome of the HYPERVIEW challenge. *IEEE Geoscience and Remote Sensing Magazine*, 12(3):35–63, sep 2024.

[7] B. G. Ram, P. Oduor, C. Igathinathane, K. Howatt, and X. Sun. A systematic review of hyperspectral imaging in precision agriculture: Analysis of its current state and future prospects. *Computers and Electronics in Agriculture*, 222:109037, 2024.

[8] F. E. Rivadeneira-Bolaños, S. E. Nope-Rodríguez, M. I. Páez-Melo, and C. R. Pinedo-Jaramillo. Automated prediction of phosphorus concentration in soils using reflectance spectroscopy and machine learning algorithms. *MethodsX*, 13:102996, 2024.

[9] B. C. Ross. Mutual information between discrete and continuous data sets. *PLOS ONE*, 9(2):e87357, 2014.

[10] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009.

[11] K. Wali, M. A. Munnaf, P. Sica, E. J. van Henten, and E. Meers. Photoacoustic fourier transform mid-infrared spectroscopy estimates plant-available phosphorus in bio-based fertilizers and amended soils. *Measurement*, 253:117651, 2025.

[12] S. Wan, J. Hou, J. Zhao, N. Clarke, C. Kempenaar, and X. Chen. Predicting soil organic matter, available nitrogen, available phosphorus and available potassium in a black soil using a nearby hyperspectral sensor system. *Sensors*, 24:2784, 2024.

[13] X. Zhang, Y. Li, and L. Zhang. Spectral-spatial graph convolutional networks for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 57(10):8481–8492, 2019.