# A Context-Aware Transformer for Soil Property Prediction via Vis-NIR Spectroscopy

Francesco De Feudis*, Ivano Bilenchi*, Agnese Pinto*, Claudio Di Paola†,
Antonio P. Leone‡, Giuseppe Loseto§, Floriano Scioscia*, Michele Ruta*
* Polytechnic University of Bari – Via E. Orabona 4, Bari (I-70125), Italy – {name.surname}@poliba.it
† MapSat S.r.l. – Contrada Santa Cumana, Benevento (I-82100), Italy – c.dipaola@mapsat.it
‡ Italian National Research Council – Piazzale E. Fermi 1, Portici (I-80055), Italy – apleone.cnr@gmail.com
§ LUM "Giuseppe Degennaro" University – Casamassima BA (I-70010), Italy – loseto@lum.it

***Abstract*—Soil characterization relies on highly accurate but costly and slow laboratory analyses. This paper presents a Transformer Encoder architecture for the simultaneous prediction of eight physico-chemical soil properties from Visible and Near-Infrared spectra. The model integrates spectral patch tokens, multi-modal context tokens encoding instrument, geographic, stratigraphic and radiometric metadata, and a global regression token. Trained and validated on the Open Soil Spectral Library dataset, the proposed model achieves $R^2 = 0.966$ for organic carbon and $R^2 = 0.918$ for $CaCO_3$.**

***Index Terms*—Soil spectroscopy, deep learning, Transformer, Vis-NIR, multi-task regression, precision agriculture, OSSL**

## I. Introduction

Physicao-chemical characterization of soils is central to sustainable agricultural resource management, fertility assessment and environmental monitoring. Traditional analytical methods, based on physical sampling and laboratory analysis, deliver high accuracy but impose substantial operational costs, slow response times and inherently limited spatial coverage [1].

Visible and Near-Infrared (Vis-NIR) spectroscopy emerges as a compelling non-destructive alternative: the spectral reflectance signature of a soil sample, acquired rapidly at low cost, encodes latent information about Organic Carbon (OC), granulometric texture, pH and Calcium Carbonate ($CaCO_3$). The relationship between spectral reflectance and pedological properties is, however, intrinsically non-linear and confounded by light-scattering effects, baseline variability and inter-laboratory artifacts. These complexities constrain traditional chemometric approaches such as Partial Least Squares Regression in cross-site and cross-instrument scenarios [1].

Recent advances in Deep Learning (DL) have demonstrated the ability of sequence models to capture long-range spectral interactions [2]. Motivated by this, the Transfomer Encoder architecture [3] – successfully applied to text (BERT [4]) and image (ViT [5]) understanding – is adapted to 1-D soil spectroscopy. The primary contributions of this research are: (i) a novel multi-modal token-encoding pipeline that jointly represents spectral patches and heterogeneous metadata; (ii) a masked multi-task loss with learned homoschedastic task weighting to handle the severe annotation sparsity of large-scale pedological datasets; (iii) a comprehensive empirical evaluation on Open Soil Spectral Library (OSSL) [6], demonstrating high predictive accuracy and lightweight deployability.

## II. Proposed Framework

### A. Dataset and Preprocessing

The OSSL Level-1 dataset [6], an open-access, globally-distributed soil spectral library, is utilized for training and evaluation. After removing samples with missing Vis-NIR spectra, $64\,384$ samples remain, each described by $1076$ reflectance bands ($350\,\mathrm{nm}$ to $2500\,\mathrm{nm}$, $2\,\mathrm{nm}$ step). Eight target properties are selected: sand, silt, clay ($\%$), pH, Electrical Conductivity (EC, $\mathrm{dS/m}$), OC ($\%$), $CaCO_3$ ($\%$) and Cation Exchange Capacity (CEC, $\mathrm{cmol/kg}$).

Spectral processing applies Standard Normal Variate (SNV) normalization sample-wise. Spectral mean ($\mu_R$) and standard deviation ($\sigma_R$), removed by SNV, are retained as the *Albedo Memory* feature vector. Textural fractions — subject to the compositional closure constraint ($\text{sand} + \text{silt} + \text{clay} = 100\,\%$) — are projected to an unconstrained Euclidean space via the Additive Log-Ratio (ALR) transformation before regression.

A site-disjoint train/validation/test split ($8 : 1 : 1$) prevents spatial data leakage.

### B. Architecture

The model input is a sequence of three token categories, described below.

**Regression token [CLS].** A learnable vector prepended to the sequence that aggregates global information through self-attention layers, following BERT [4] and ViT [5].

**Context tokens.** Seven tokens encode exogenous metadata: four categorical instrument tokens [INS] (spectrometer model, device code, optical configuration, sample preparation protocol) as learned embeddings; a depth token [DEP] (upper/lower horizon bounds); and an albedo token [ALB] $(\mu_R, \sigma_R)$. Continuous tokens are projected via dedicated two-layer feed-forward networks. Missing context tokens are excluded from self-attention via binary padding masks, preserving partially available information.

**Spectral tokens.** The SNV-normalized spectrum is encoded by a 1-D convolutional *latent encoder* with kernel width $w$, stride $s$ and $d$ filters, producing $n \approx L/s$ patch embeddings followed by Sigmoid-weighted Linear Unit (SiLU) activation

TABLE I
PREDICTIVE METRICS ON THE TEST SET.

| Property | MAE | $R^2$ | RPD | RPIQ |
|---|---|---|---|---|
| Sand (%) | 9.92 | 0.704 | 1.84 | 3.05 |
| Silt (%) | 7.57 | 0.656 | 1.70 | 2.51 |
| Clay (%) | 4.34 | 0.787 | 2.17 | 2.76 |
| pH | 0.33 | 0.886 | 2.97 | 5.39 |
| EC (dS/m) | 0.11 | 0.520 | 1.44 | 0.87 |
| OC (%) | 1.08 | **0.966** | **5.46** | 1.54 |
| $CaCO_3$ (%) | 1.36 | 0.918 | 3.50 | 0.58 |
| CEC (cmol/kg) | 3.74 | 0.823 | 2.37 | 2.12 |

[7] and Layer Normalization [8]. Sinusoidal positional encoding is added to inject wavelength ordering [3].

The full sequence is processed by a standard Transformer Encoder [3] with multi-head self-attention. The final [CLS] representation is fed to $T = 7$ independent regression heads (one per target scalar property in Table I, except Clay which is obtained from the compositional closure constraint), each a two-layer feed-forward with Gaussian Error Linear Unit (GELU) [9] activation and head-specific dropout.

### C. Training Objective

A masked multi-task Huber loss handles missing annotations: per-task losses are computed only on valid samples. Automatic task weighting follows [10]: each task $\tau$ has a learnable log-uncertainty $\log \omega_\tau$, and the aggregated loss is:

$$\mathcal{L} = \sum_{\tau} \left[ e^{-\log \omega_\tau} \cdot \mathcal{L}_\tau + \log \omega_\tau \right] \quad (1)$$

where $\mathcal{L}_\tau$ denotes the Huber loss [11]. Tasks the model deems inherently noisy receive smaller gradients, preventing unreliable targets from corrupting the shared latent representation.

## III. Early Evaluation

### A. Hyperparameter Optimization

Bayesian optimization with Hyperband early stopping [12] selects the configuration maximizing validation performance: $d = 64$, 8 attention heads, 7 Transformer layers, $m_{ff} = 2$, $w = 64$, and $s = 32$ (overlap $o = 0.5$). The final model comprises $253\,415$ parameters ($3.37\,\mathrm{MB}$).

### B. Regression Performance

Table I reports the predictive metrics on the held-out test set. Performance is heterogeneous, reflecting the differing spectral expressiveness of each target.

OC ($R^2 = 0.966$, $\mathrm{RPD} = 5.46$) and $CaCO_3$ ($R^2 = 0.918$, $\mathrm{RPD} = 3.50$) benefit from characteristic Near-Infrared (NIR) absorptions and high annotation density. pH ($R^2 = 0.886$) and CEC ($R^2 = 0.823$) comfortably exceed the $\mathrm{RPD} \geq 2.0$ threshold. Textural fractions reach acceptable accuracy ($\mathrm{RPD} \in [1.70, 2.17]$); EC is the most challenging parameter ($R^2 = 0.520$), consistent with the literature on its limited Vis-NIR spectral signature [1].

The learned homoschedastic task weights closely mirror the empirical rankings, confirming that the adaptive loss correctly suppresses noisy gradients from uncertain targets and focuses capacity on the most spectrally expressive properties.

## IV. Conclusion

The proposed Transformer Encoder model, augmented with multi-modal context tokens and uncertainty-aware multi-task training, can simultaneously infer eight soil physico-chemical properties from Vis-NIR spectra with accuracy competitive or superior to specialized chemometric models. Its 253K-parameter design couples high predictive performance with extreme computational efficiency, enabling real-time *in situ* inference. Future work will include self-supervised spectral pre-training on unlabeled data, ablation studies to quantify the individual contribution of each context modality, and experimental comparisons with both a chemometric and a deep-learning baseline on the same data and metrics.

## Acknowledgments

This work has been supported by the “Ap-EDIH – Apulian European Digital Innovation Hub” project, funded under the NextGenerationEU program by the Italian Ministry of Enterprises and Made in Italy, CUP B97H22005100008.

## References

[1] R. V. Rossel, D. Walvoort, A. McBratney, L. J. Janik, and J. Skjemstad, “Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties,” *Geoderma*, vol. 131, no. 1-2, pp. 59–75, 2006.

[2] H. Wang, Q. Sun, X. Niu, K. Liu, J. Zhang, Z. Hao, and D. Xu, “Soil Organic Carbon Prediction Using an Efficient Channel Attention-Enhanced CNN-LSTM Model With LUCAS Spectral Library,” *European Journal of Soil Science*, vol. 76, no. 5, p. e70202, 2025.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” in *International Conference on Learning Representations*, 2021.

[6] J. L. Safanelli, T. Hengl, L. L. Parente, R. Minarik, D. E. Bloom, K. Todd-Brown, A. Gholizadeh, W. d. S. Mendes, and J. Sanderman, “Open Soil Spectral Library (OSSL): Building reproducible soil calibration models through open development and community engagement,” *PloS one*, vol. 20, no. 1, p. e0296545, 2025.

[7] S. Elfwing, E. Uchibe, and K. Doya, “Sigmoid-weighted linear units for neural network function approximation in reinforcement learning,” *Neural networks*, vol. 107, pp. 3–11, 2018.

[8] J. L. Ba, J. R. Kiros, and G. E. Hinton, “Layer normalization,” *arXiv preprint arXiv:1607.06450*, 2016.

[9] D. Hendrycks and K. Gimpel, “Gaussian error linear units (gelus),” *arXiv preprint arXiv:1606.08415*, 2016.

[10] A. Kendall, Y. Gal, and R. Cipolla, “Multi-task learning using uncertainty to weigh losses for scene geometry and semantics,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7482–7491.

[11] P. J. Huber, “Robust estimation of a location parameter,” in *Breakthroughs in statistics: Methodology and distribution*. Springer, 1992, pp. 492–518.

[12] S. Falkner, A. Klein, and F. Hutter, “BOHB: Robust and efficient hyperparameter optimization at scale,” in *International conference on machine learning*. PMLR, 2018, pp. 1437–1446.