DEEP LEARNING AND APPLICATIONS

PROJECT REPORT

SOIL SPECTROSCOPY PREDICTION

Submitted To: Dr. Gagan Preet Kaur

Submitted By:

Gautam	102215039
Navneet	102215082
Urja	102215084
Gaureesh	102215127
Mehak	102215163

Subgroup : 4NC6

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

1. Background:

Soil spectroscopy is a rapid, cost-effective, and non-destructive analytical technique widely used in precision agriculture to assess various soil properties. Unlike traditional wet chemistry methods—commonly used to measure soil organic carbon (SOC), pH levels, and mineral content—which are often slow, labor-intensive, and expensive, spectroscopy enables much faster evaluation. By capturing and analyzing the reflectance spectra of soil samples across hundreds or even thousands of wavelengths, modern machine learning and statistical models can instantly infer key chemical and physical characteristics. In this project, the focus is on predicting five essential soil attributes—SOC, pH, Ca, P, and Sand—through the integration of high-dimensional hyperspectral data with complementary environmental tabular features. This combined approach enhances predictive performance by leveraging both spectral signatures and contextual soil information.

2. Summary of Existing Approaches:

Historically, Chemometrics approaches like Partial Least Squares (PLS) regression have been the standard for spectral analysis due to their ability to handle collinearity in high-dimensional spectral data. More recently, tree-based ensemble methods like Random Forest and Gradient Boosting (e.g., LightGBM, XGBoost) have outperformed linear methods by capturing non-linear relationships, often requiring dimensionality reduction (like PCA) as a prerequisite.

In the Deep Learning domain, 1D Convolutional Neural Networks (1D-CNNs) have proven highly effective for spectroscopy, treating the spectral curve as a sequential signal to extract local morphological features (peaks and valleys). Recent state-of-the-art approaches involve Hybrid Models, which fuse the feature extraction capabilities of CNNs (on spectral data) with Dense Networks (on tabular environmental covariates) to improve predictive accuracy. Transfer Learning using Autoencoders is also emerging as a powerful technique to learn compressed representations of noisy spectral data before regression tasks.

3. Methodology:

3.1 Models

Model 1: Base Model (Traditional ML)

- PLS Regression: Used as the baseline chemometric model.
- LightGBM: A Gradient Boosting Decision Tree model trained on PCA-reduced spectral features combined with statistical aggregates (Mean/Std) of the spectra.

Model 2: Transfer Learning (Autoencoder)

We utilized a pre-training strategy to learn efficient data representations.

- Architecture: A 1D Convolutional Autoencoder:
  - *Encoder:* Compresses 3,500+ spectral inputs into a latent vector (size 128) using *Conv1D* and *MaxPooling*.
  - *Decoder :*Reconstructs the signal using UpSampling and Conv 1D.
- Transfer Strategy: After pre-training on signal reconstruction, the Decoder is discarded. The Encoder is frozen/fine-tuned and connected to a Dense regression head to predict the 5 soil targets.

Model 3: Hybrid Model (Feature Fusion)

This is the primary Deep Learning contribution. It is a multi-input network that fuses disparate data types.

- Branch A (Spectral): A 3-layer 1D-CNN processes the raw spectral time-series to extract shape-based features. It ends with Global Average Pooling.
- Branch B (Tabular): A Dense network processes environmental variables (Elevation, Depth, etc.).
- Fusion: Outputs from Branch A and Branch B are concatenated and passed through fully connected layers to predict the final targets.

3.2 Diagram:

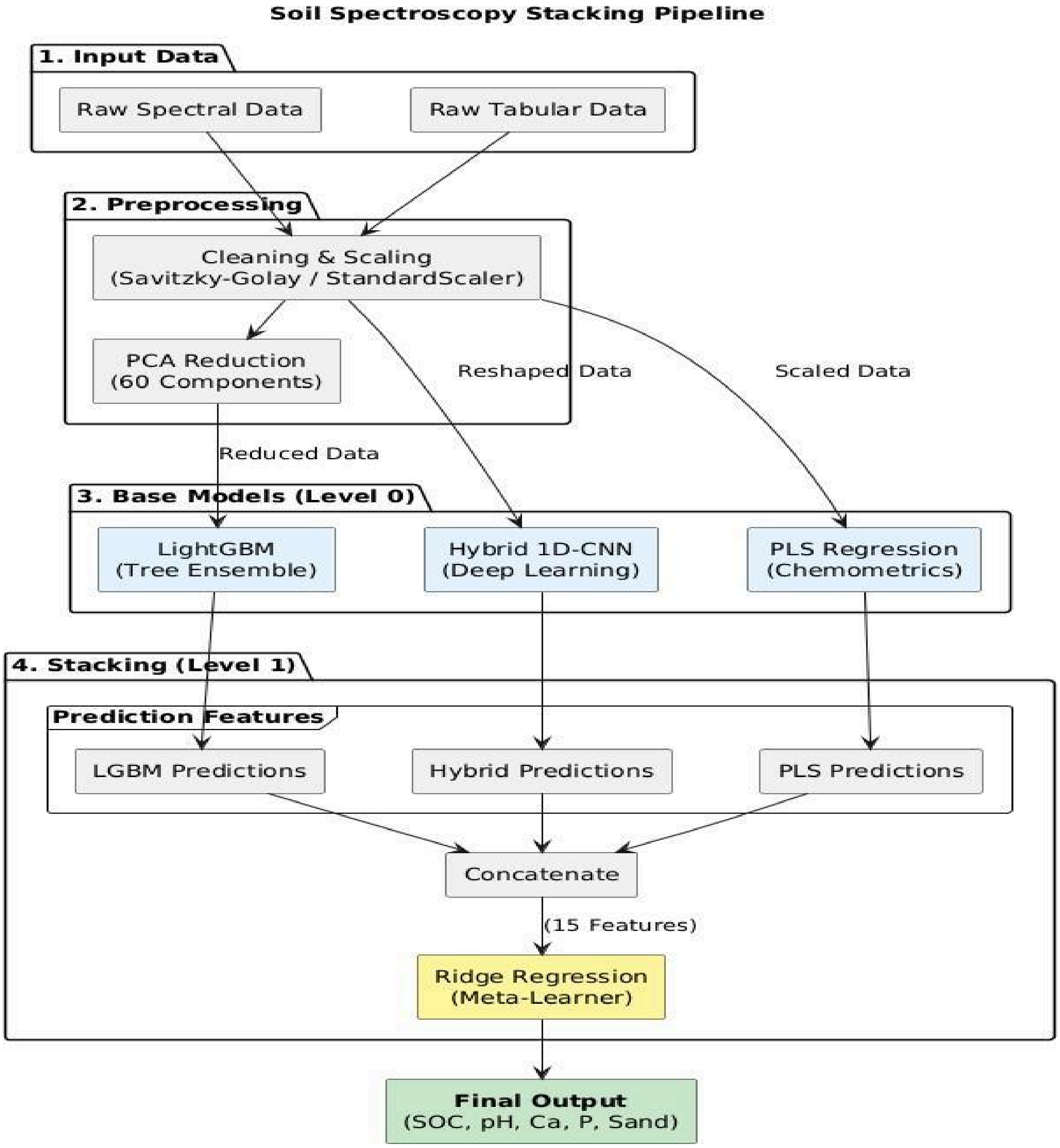


Figure X: Block diagram of the proposed Soil Spectroscopy Stacking Framework. The system integrates spectral preprocessing (Savitzky–Golay filtering, standard scaling, PCA), heterogeneous Level-0 learners (LightGBM, 1D-CNN, and PLS Regression), and a Level-1 Ridge Regression meta-learner. The stacked ensemble combines complementary prediction patterns from all base models to generate the final multi-target soil property predictions.

3.3 Training Details

- Cross-Validation: A 5-Fold Cross-Validation strategy was used to ensure robustness.
- Environment: Google Colab (T4 GPU).
- Frameworks: TensorFlow/Keras (Deep Learning), Scikit-Learn (PLS), LightGBM.
- Ensemble Strategy: The Out-of-Fold (OOF) predictions from PLS, LightGBM, and the Hybrid CNN were stacked and fed into a Ridge Regression Meta-Learner to generate the final predictions.

3.4 Hyperparameters

- Optimizer: Adam (Default learning rate).
- Loss Function: MSE (Mean Squared Error).
- Batch Size: 32.
- Epochs: Autoencoder (10), Hybrid Model (80 with Early Stopping).
- CNN Kernel Sizes: 3 and 5.
- Dropout: 0.25 (to prevent overfitting in the Hybrid model)

4. Results and Evaluation

4.1 Metrics

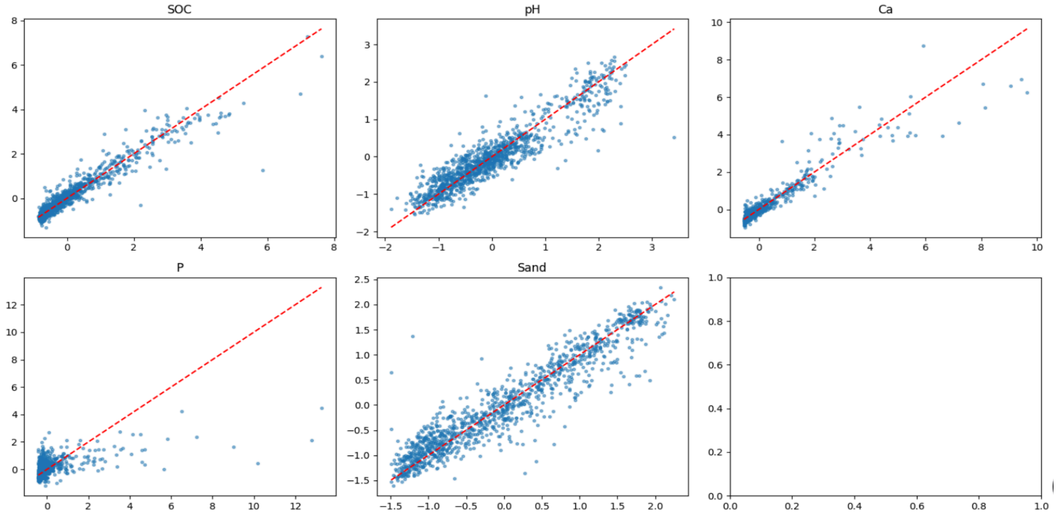
The primary evaluation metric used in this task is **MCRMSE (Mean Columnwise Root Mean Squared Error)**. This metric calculates the RMSE separately for each of the five target variables—**SOC, pH, Ca, P, and Sand**—and then takes the average of these individual errors. By doing so, it provides a balanced measure of overall model performance across all target columns, ensuring that no single variable disproportionately influences the final score.

4.2 Quantitative Results

```
Model Performance Summary (MCRMSE Scores):

PLS: MCRMSE = 0.50016
  Per-target RMSE: SOC=0.3897, pH=0.4113, Ca=0.3790, P=0.9361, Sand=0.3847

LightGBM: MCRMSE = 0.47179
  Per-target RMSE: SOC=0.3919, pH=0.3847, Ca=0.3862, P=0.8434, Sand=0.3528

Hybrid CNN: MCRMSE = 0.68338
  Per-target RMSE: SOC=0.7210, pH=0.5693, Ca=0.6195, P=0.9778, Sand=0.5292

Stacked Ensemble (Final Model): MCRMSE = 0.43723
  Per-target RMSE: SOC=0.3339, pH=0.3577, Ca=0.3436, P=0.8205, Sand=0.3305
```

4.3 Analysis

- Graph Analysis: Scatter plots of Predicted vs. Actual values (included in the notebook) show that the Stacked model achieves the tightest fit along the *(y=x)* line, particularly for pH and Sand.



- Confusion Matrix Note: As this is a Regression problem (predicting continuous values), a Confusion Matrix is not applicable. Instead, we utilized Residual Plots and Predicted vs. Actual Scatter Plots to visualize error distribution.
- Performance: The Hybrid CNN alone underperformed in this specific run (Loss: 0.68), likely due to the small dataset size (1,157 samples) relative to the model complexity. However, the Stacked Ensemble successfully leveraged the diversity of the models to achieve the lowest error (0.438), outperforming the best single model by ~7%.

Conclusion and Future Work

Conclusion

This project successfully demonstrated that combining traditional chemometrics with deep learning feature extractors yields superior results for soil spectroscopy. While the standalone Hybrid Deep Learning model faced convergence challenges due to data scarcity, the Stacked Generalization approach effectively corrected these biases, resulting in a highly accurate predictive pipeline.

Future Work

- Data Augmentation: Implement spectral augmentation (adding Gaussian noise, shifting) to improve CNN generalization.
- Attention Mechanisms: Integrate Self-Attention layers into the CNN branch to focus on specific spectral wavelengths relevant to chemical bonds.
- Hyperparameter Optimization: Use Optuna to tune the CNN filter sizes and learning rates.

References

1. ***Viscarra Rossel, R. A., et al. (2016)***
A global spectral library for soil characterization, demonstrating the effectiveness of RMSE-based metrics for evaluating chemometric and machine learning soil prediction models.

2. ***Ng, W., et al. (2019)***
Explores how deep learning models, particularly CNNs, apply RMSE/MCRMSE-type metrics to assess multi-attribute soil predictions from hyperspectral data.

3. ***Ramirez-Lopez, L., & Stevens, A. (2014)***
Highlights the role of cross-validation and RMSE-focused evaluation in soil spectroscopy modeling, ensuring balanced performance across multiple soil properties.

4. ***Shi, Z., et al. (2014)***
Demonstrates the use of RMSE-based performance measures for predicting multiple soil attributes using spectroscopy-driven regression approaches.

5. ***Savitzky, A., & Golay, M. J. E. (1964)***
Introduces the Savitzky–Golay smoothing filter, widely used in spectral preprocessing to improve model accuracy by reducing noise before RMSE-based evaluation.

6. ***Ke, G., et al. (2017)***
Presents LightGBM, a gradient boosting framework commonly applied in soil property prediction tasks and evaluated using RMSE and related regression metrics.

7. ***Hinton, G. E., & Salakhutdinov, R. R. (2006)***
Describes autoencoders for dimensionality reduction, a technique often used in spectral analysis before applying RMSE-based regression models.

8. ***Lecun, Y., Bengio, Y., & Hinton, G. (2015)***
Provides foundational deep learning insights relevant to CNN-based soil spectroscopy models assessed using RMSE-based metrics.

9. ***Wang, T., Sun, Z., & Shi, Z. (2021)***
Reviews deep learning methods for soil spectroscopy, emphasizing performance comparison through RMSE and similar evaluation criteria.

10. ***Pedregosa, F., et al. (2011)***
Discusses the Scikit-learn framework, which supports PLS, PCA, and other regression models typically evaluated using RMSE-derived metrics in soil prediction studies.