Development of a Gaussian Process Regression Model for Soil Composition Prediction Using Multispectral Image Analysis

A thesis submitted in partial fulfillment of the requirements for
the award of the degree of

B.Tech

in

Electrical and Electronics Engineering

By

Harshul D (107121041)

Pusala Surendra Babu (107121079)

ELECTRICAL AND ELECTRONICS ENGINEERING
NATIONAL INSTITUTE OF TECHNOLOGY

TIRUCHIRAPPALLI – 620015

JUNE-JULY 2024

ABSTRACT

This report presents the methodology, and results of our internship project focused on predicting the proportions of sand, silt, and clay in soil samples using multispectral images captured by the Parrot Sequoia camera. The project addresses the challenge of transforming low-resolution multispectral data into a format suitable for regression analysis. The methodology involved data collection, image preprocessing, and regression modelling using Gaussian Process Regression (GPR). Key steps included cropping raw images, separating RGB bands, and subblocking images to extract essential features. We then formulated the Gram matrix and applied the Fast PCA technique for dimensionality reduction and data normalization. A target matrix was developed to map image features to soil composition ratios, with GPR regression applied to predict the proportions. The results demonstrate that our model can effectively predict soil composition, offering a promising approach for precision agriculture and environmental monitoring. Further research may focus on enhancing model accuracy through additional data or more advanced techniques.

Keywords : Fast Principal Component Analysis, Gaussian Process Regression.

ACKNOWLEDGEMENTS

We would like to express our deepest gratitude to our internship advisor for guiding us through this course and without whom this project and the results achieved from it would not have reached completion.

Dr. E.S. Gopi, Associate Professor, Department of Electronics and Communication Engineering, for helping us and guiding us in the course of this project. Without his guidance, we would not have been able to successfully complete this project. His patience and genial attitude is and always will be a source of inspiration to us.

We are also thankful to the Ms. Vinodha, a PhD scholar from the Department of Electrical and Electronics Engineering for her assistance in the intership, our individual parents and our friends for their constant support and help.

TABLE OF CONTENTS

Chapter 5 : Conclusion

Chapter 1: Introduction

1.1 Background

Soil composition plays a critical role in determining the fertility, structure, and overall health of agricultural fields. Understanding the proportions of sand, silt, and clay within soil is fundamental for optimizing crop yield and managing land resources. Traditional soil analysis methods, while effective, are often time-consuming and require extensive manual sampling. With advances in imaging technologies and data analytics, multispectral imaging offers a non-invasive alternative for soil analysis. This project leverages multispectral data captured by the Parrot Sequoia camera to predict soil composition, aiming to improve the efficiency and accuracy of soil assessment for precision agriculture.

1.2 Objectives

The primary objective of this project is to develop a regression model capable of predicting the proportions of sand, silt, and clay in soil samples based on multispectral images. The specific objectives are:

- To collect and preprocess multispectral images of soil samples.
- To extract relevant features from these images for regression analysis.
- To apply dimensionality reduction techniques for enhancing model performance.
- To develop a Gaussian Process Regression (GPR) model for accurate prediction of soil composition.

1.3 Problem Statement

Current soil composition analysis methods rely heavily on physical sampling and laboratory testing, which are labor-intensive and time-consuming. With growing demands for real-time data in precision agriculture, there is a need for efficient and scalable techniques that provide accurate soil composition estimates using imaging technologies. This project addresses the challenge of predicting soil composition proportions from multispectral images, aiming to bridge the gap between traditional methods and advanced imaging-based analysis.

1.4 Contributions

This project contributes to the field of soil analysis by:

- Proposing an automated approach to crop and preprocess multispectral images for soil composition prediction.
- Developing a robust feature extraction pipeline involving Gram matrix formulation and Principal Component Analysis (PCA).
- Introducing a GPR-based regression model tailored for soil composition prediction.
- Demonstrating the practical applicability of the model in real-world scenarios, particularly for precision agriculture.

Chapter 2: Literature Review

2.1 Overview of Multispectral Imaging in Soil Analysis

Multispectral imaging has emerged as a powerful tool in precision agriculture, enabling non-invasive analysis of soil properties through data captured across different spectral bands. Various studies have explored its potential in assessing soil characteristics, such as moisture content, organic matter, and texture. The use of multispectral imaging allows for the capture of subtle differences in reflectance that correlate with soil composition, providing a robust foundation for predictive modeling. Researchers have highlighted the importance of selecting the right spectral bands and preprocessing techniques to enhance model accuracy, particularly when predicting physical properties like sand, silt, and clay ratios.

2.2 Soil Composition Prediction Techniques

Traditional methods for soil composition analysis involve physical sampling followed by laboratory tests, which are accurate but time-intensive. In recent years, researchers have explored regression-based models that use imaging data as a more efficient alternative. Multiple studies have focused on predicting soil texture using various types of imagery, including multispectral and hyperspectral data. Among the approaches explored, regression models like Support Vector Regression (SVR), Random Forest Regression, and Partial Least Squares Regression (PLSR) have been widely applied. These methods, however, often require extensive feature engineering and large datasets, limiting their scalability in real-world applications. Our work aims to leverage these insights while incorporating advanced techniques like Gaussian Process Regression (GPR) to improve prediction accuracy.

2.3 Feature Extraction and Dimensionality Reduction in Image Analysis

Feature extraction is crucial when working with multispectral images, as it determines the quality of the input data for regression models. The Gram matrix formulation, which captures relationships between different spectral bands, has been proposed as an effective feature extraction technique in recent studies. Dimensionality reduction methods like Principal Component Analysis (PCA) are often employed to minimize computational complexity while retaining essential data features. Research shows that combining these approaches enhances the performance of predictive models by reducing noise and redundancy in high-dimensional data.

2.4 Gaussian Process Regression (GPR) in Predictive Modelling

Gaussian Process Regression (GPR) has gained attention for its flexibility in handling non-linear and sparse datasets, making it particularly suitable for soil analysis where data variability is common. Unlike other regression models, GPR provides probabilistic predictions, offering uncertainty estimates that can be valuable for decision-making in agricultural applications. Studies have applied GPR successfully in scenarios involving environmental monitoring, remote sensing, and geospatial analysis, demonstrating its robustness in dealing with complex, high-dimensional datasets. By incorporating GPR into our methodology, we aim to leverage these advantages to produce reliable predictions of soil composition proportions from multispectral images.

2.5 Challenges and Gaps in Existing Research

Despite significant advancements, challenges remain in applying multispectral imaging and regression models for soil analysis. One of the primary issues is the variability in soil properties across different geographic regions, which limits model generalization. Additionally, multispectral data often suffers from noise and overlapping spectral information, leading to inaccuracies in feature extraction. The literature also points to a lack of standardized preprocessing workflows for image-based soil analysis, making it difficult to replicate results across different studies. Our project addresses these gaps by focusing on a systematic approach to image preprocessing, feature extraction, and regression modeling that aims to be both scalable and adaptable across various soil types.

2.6 Summary and Positioning of Our Work

This literature review highlights the potential of multispectral imaging combined with advanced regression techniques for soil composition prediction. While existing studies provide a strong foundation, they also reveal gaps in terms of model robustness, feature extraction, and generalization across different environments. Our work builds on these findings by proposing a comprehensive pipeline that integrates cutting-edge techniques like Gram matrix formulation, PCA for dimensionality reduction, and GPR for predictive modeling. By focusing on these areas, our project contributes a novel approach that enhances the accuracy and applicability of soil composition analysis in precision agriculture.

Chapter 3: Methodology

3.1 Soil Image Acquisition and Preprocessing

Accurate soil composition prediction begins with the acquisition and preprocessing of multispectral images. In this project, soil samples were captured using the Parrot Sequoia camera, which provides high-resolution multispectral images across five bands: green, red, red edge, and near-infrared (NIR), along with an RGB image. In total, the dataset comprises 186 samples, collected as 3 distinct samples for each of the 62 different soil proportions, with each sample captured across 5 spectral bands. These images form the basis for subsequent feature extraction and analysis. The preprocessing stage involved two key steps: image cropping and RGB channel separation.

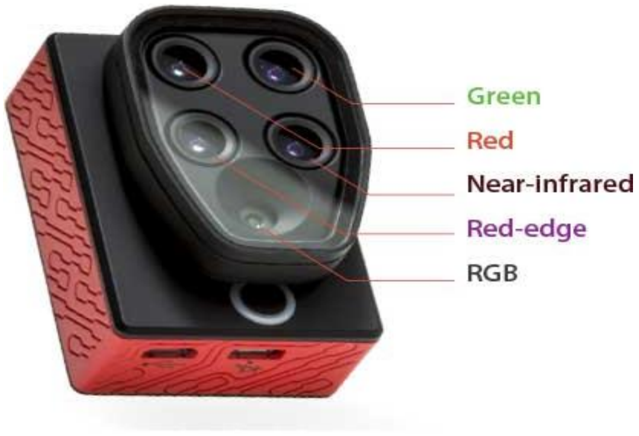
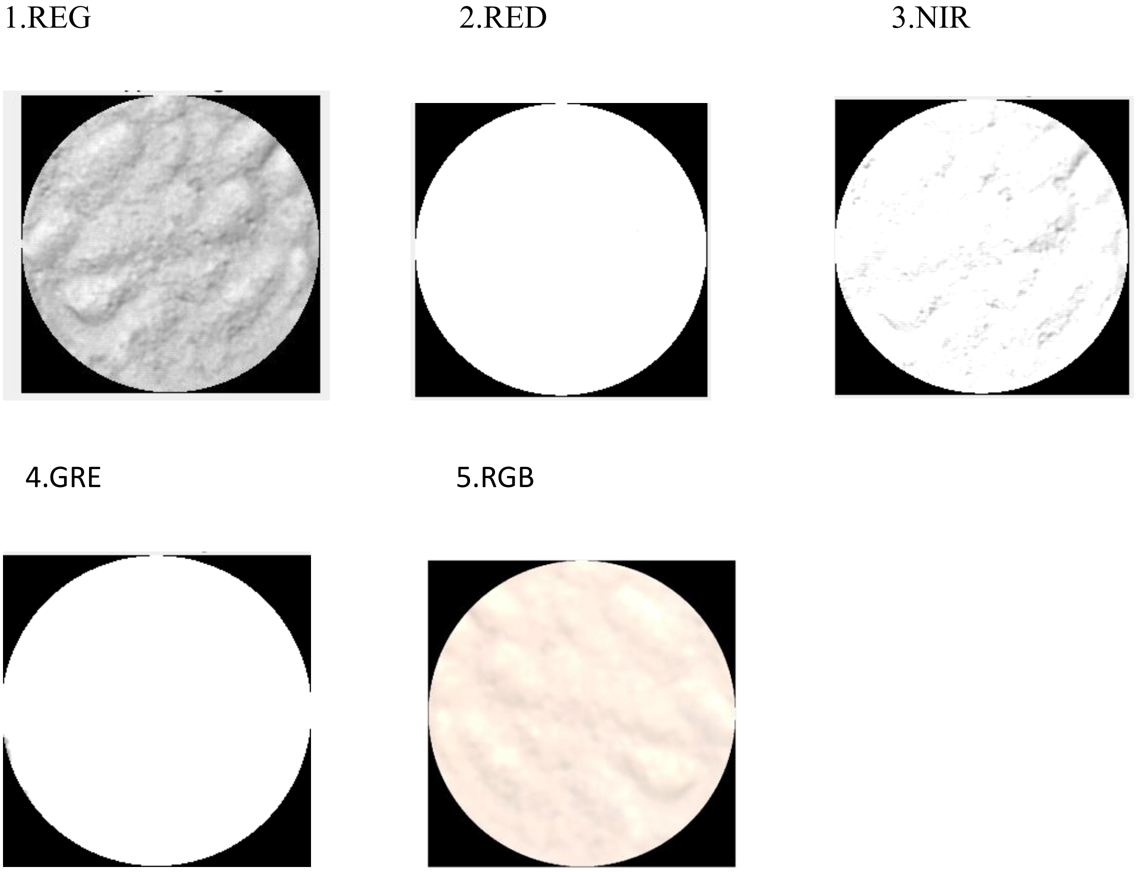


Figure 1: PARROT SEQUOIA + MULTISPECTRAL CAMERA

3.1.1 Image Cropping

The initial images contained significant background noise and extraneous details unrelated to the soil samples. To isolate the relevant regions, we developed a cropping routine in MATLAB. This routine identified the portion of the image containing the soil sample and removed unnecessary borders and artifacts. Given that the RGB image was rotated 180 degrees compared to the other bands, a correction was applied to ensure alignment. The cropped images were then standardized to a consistent size, allowing for uniform processing in subsequent stages.



1.REG 2.RED 3.NIR

4.GRE 5.RGB

Figure 2: Cropped Images

3.1.2 RGB Channel Separation

Following image cropping, the RGB image was split into its constituent Red, Green, and Blue channels. Each channel was treated as a distinct input band, expanding the available spectral data for analysis. This step was essential for enhancing feature extraction by allowing the model to treat each channel as an independent source of information. The separation process was applied consistently across all samples, ensuring that each image was correctly decomposed into its respective channels before being integrated with data from the other multispectral bands.

3.2 Subblocking and Feature Extraction

After preprocessing the images, the next step involves dividing the images into smaller, more manageable sections (subblocks) and extracting relevant features that can be used for predictive modeling. This stage is crucial for capturing localized information from different regions of the soil samples, which can provide more detailed insights into their composition.

3.2.1 Dividing Images into Subblocks

To effectively capture localized features, each cropped image was divided into 10 subblocks, each of size 100x100 pixels. The subblocking process was designed to focus on key regions within the image while maintaining overlap between adjacent subblocks to enhance feature continuity. The overlap ratio was set at 0.9, ensuring that subblocks share significant portions of data, which helps in better capturing transitions and gradients within the soil sample.

The subblocking was implemented using a systematic method that starts from the center of the image and expands outward. The process identifies the central point of the image and iteratively defines subblocks by moving outward in a grid pattern. The overlap between subblocks is carefully maintained to avoid loss of critical information and ensure consistency across all images. This method prioritizes the central region, which typically contains the most representative data, while minimizing noise that might be introduced from the edges.

This approach was chosen to ensure that the most relevant regions of the image are captured while minimizing edge effects that could introduce inconsistencies in feature extraction. By using subblocks with significant overlap, the model can better capture local variations in soil composition, leading to more accurate predictions.

3.2.2 Gram Matrix Construction

The Gram matrix in this context is a feature matrix that organizes the spectral data for each subblock. For each subblock, the 7 bands are flattened into 10,000-element vectors (from their original 100x100 pixel grids) and stacked column-wise, resulting in a 10000 x 7 matrix.

Given that our dataset consists of 62 distinct soil types, each with 3 samples, the overall dataset contains 186 samples, yielding a total of 1,840 subblocks (With 20 missing subblocks). When organized collectively, this results in a final Gram matrix of size 70000 x 1840 (70,000 rows representing pixels across each of the subblocks, and 1,840 columns representing each subblock across all samples). This matrix structure encapsulates the critical spectral data needed for the regression model, providing a compact yet comprehensive representation of the soil samples for further analysis.

3.3 Dimensionality Reduction using Fast PCA

In high-dimensional datasets, reducing the number of features is essential to minimize computational complexity while retaining the most significant information. In this project, Principal Component Analysis (PCA) was employed to achieve dimensionality reduction, with a focus on optimizing the process using Fast PCA.

3.3.1 Principal Component Analysis (PCA) Overview

Principal Component Analysis is a widely used technique in data processing that projects high-dimensional data onto a lower-dimensional space. The core idea is to identify directions (principal components) that capture the maximum variance in the data. Mathematically, this is achieved by performing an eigen decomposition on the covariance matrix of the data.

For a given data matrix X of size m×n (where n represents the number of samples and m the number of features), the covariance matrix Σ is defined as:

$$\varSigma = \frac{1}{m} X^T \cdot X$$

The eigenvectors of Σ correspond to the principal components, and the eigenvalues indicate the amount of variance captured by each component. The principal components with the largest eigenvalues are selected, allowing the data to be projected onto a subspace that retains the most relevant information.

3.3.2 Application of Fast PCA

Given the large size of our Gram matrix (70,000 × 1,840), performing traditional PCA would be computationally intensive. We employed Fast PCA, which optimizes the process by working within a reduced space. The approach begins by calculating the inner product matrix $I = M^T . M$, resulting in a more manageable 1,840 × 1,840 matrix. This smaller matrix captures the variance structure across samples, making it computationally feasible.

Next, we compute the covariance matrix after centering I by subtracting the mean of each column. To manage numerical stability, particularly in cases where I isn't full-rank, we compute the pseudoinverse. We then solve the eigenvalue problem for $I_{pinv}.cov(I)$ to extract eigenvectors (principal components) and eigenvalues (variance captured by each component). The top eigenvectors, which capture a significant percentage of the variance (80% in our case) are selected.

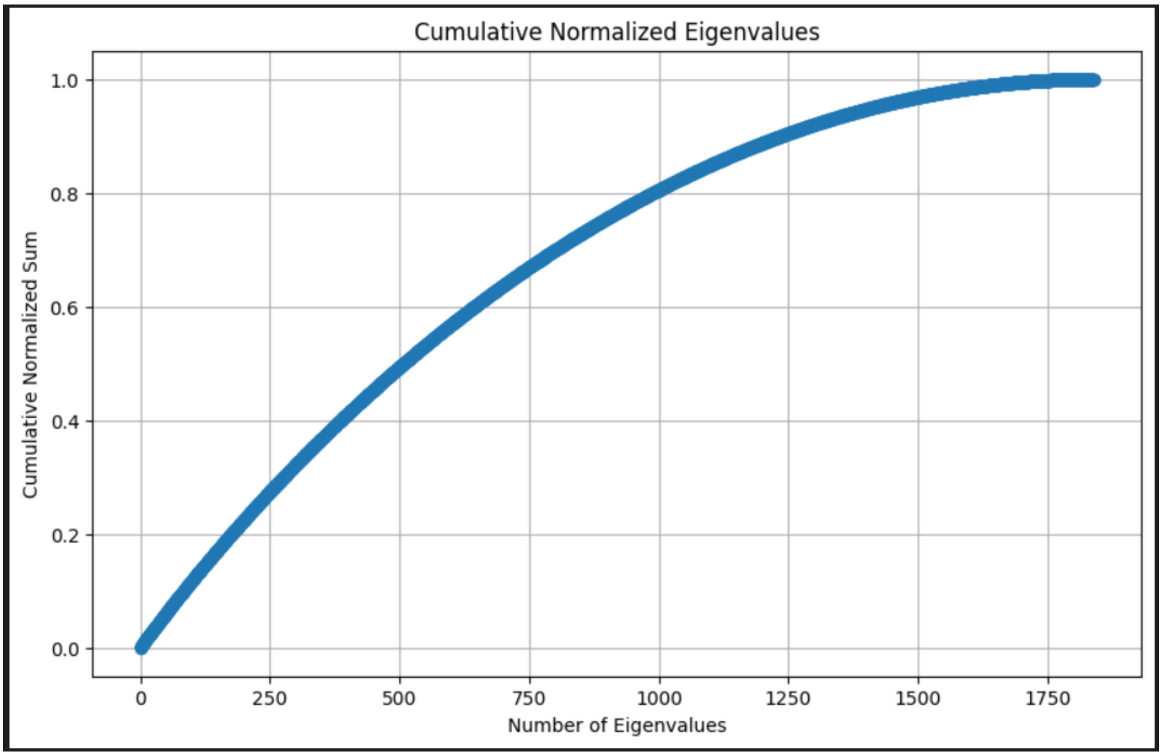


Figure 3 : Cumulative Normalized EigenValues

These selected eigenvectors form the basis for projecting the original data into a lower-dimensional space. The projection results in a reduced data matrix Z, with dimensions k×1840 where k represents the number of principal components chosen (k = 990 in our case). This reduced matrix retains the most significant information while allowing for efficient modeling. This reduced data matrix is further normalized to ensure that all features contribute equally in subsequent modeling.

3.4 Target Matrix Formulation

The target matrix plays a crucial role in linking the extracted features from the Gram matrix to the actual soil composition proportions (sand, silt, and clay) that we aim to predict. Each soil sample in our dataset is associated with known proportions of these three components, and this information forms the basis of the target matrix.

Thus we will obtain a target matrix of size 3 x 1840, with the rows corresponding to the three target variables and columns corresponding to the number of image samples. This arrangement is critical for supervised learning, allowing the regression model to learn the relationships between the features (input data) and the soil composition proportions (output data).

3.5 Predictive Modeling with Gaussian Process Regression

Gaussian Process Regression (GPR) was chosen as the predictive modeling technique due to its ability to model complex, non-linear relationships and provide uncertainty estimates alongside predictions. This flexibility is particularly advantageous for our dataset, where the relationships between the spectral features and soil composition proportions are intricate and not easily captured by simpler regression methods.

3.5.1 Gaussian Process Regression (GPR) Model Development

GPR is a non-parametric, probabilistic model that assumes a prior distribution over functions and updates this distribution based on observed data. The core of GPR lies in its kernel function, which determines the similarity between data points and thus influences the predictions. For this project, the Radial Basis Function (RBF) kernel was employed, which is effective in capturing smooth, non-linear patterns. The RBF kernel is defined as:

$$\mathrm{k}(\mathrm{x}_{\mathrm{i}}\,,\mathrm{x}_{\mathrm{j}}) = exp\left(\frac{\|x_i - x_j\|}{2\sigma^2}\right)$$

where x_i and x_j are feature vectors, and σ controls the width of the kernel. This kernel choice allows the model to adapt to varying scales of data and capture the relationships that are most relevant for predicting soil composition.

3.5.2 Training and Testing

The model was trained using the reduced and normalized feature set obtained from Fast PCA. The target matrix, containing the proportions of sand, silt, and clay, was used as the output variable. A 80-20 split was applied to the dataset, where 80% of the samples were used for training and the remaining 20% for testing. The training process involved optimizing the hyperparameters of the GPR model, such as the length scale and noise level, to minimize prediction error.

During testing, the model’s performance was evaluated using Mean Squared Error (MSE). MSE measures the average squared difference between predicted and actual values, with lower MSE values indicating better model performance. By minimizing MSE, we ensure that the model’s predictions closely align with the actual soil composition proportions, reflecting the quality and reliability of the regression model.

Chapter 4: Results and Discussion

4.1 Model Performance Evaluation

The performance of the Gaussian Process Regression (GPR) model was evaluated using the Mean Squared Error (MSE), which was found to be 0.0049. This low MSE value indicates that the model achieved high accuracy in predicting the proportions of sand, silt, and clay. The model successfully captured the complex relationships in the spectral data, leading to predictions that align closely with the actual values.

4.2 Predicted vs. Actual Comparisons

The figures below show the comparisons between predicted and actual values for sand, silt, and clay. As seen in the plots, the predicted values (orange lines) closely follow the true values (blue lines), demonstrating that the model performs well across the dataset. Although there are slight variations, the overall alignment suggests that the model is capable of making reliable predictions for soil composition.

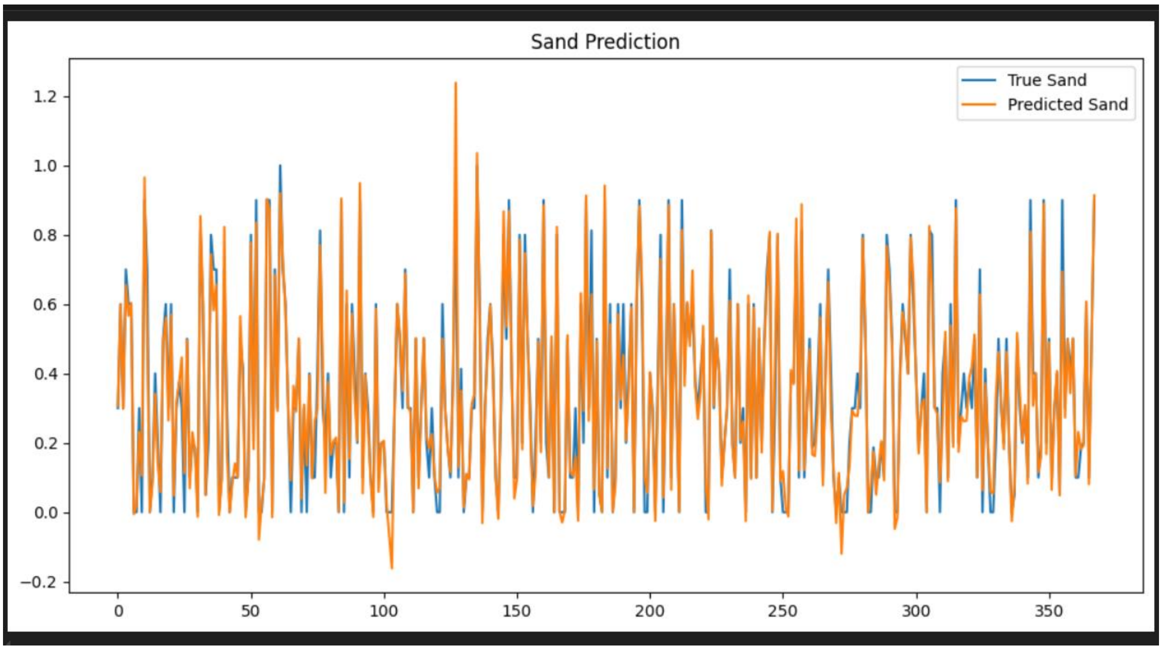


Figure 4 : Sand Prediction

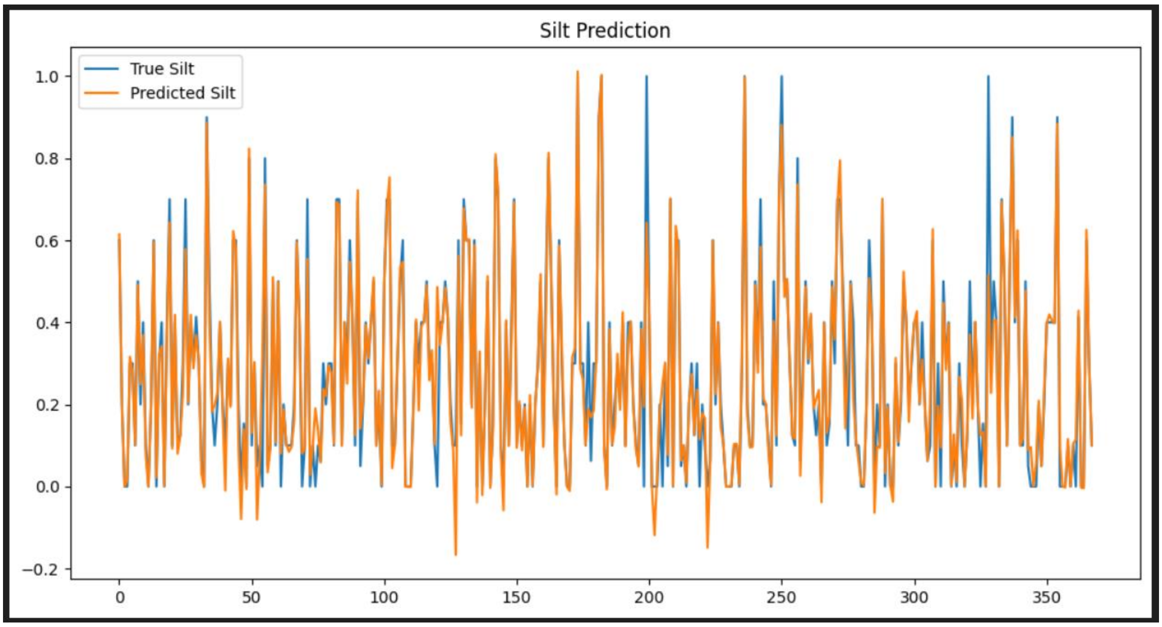


Figure 5 : Slit Prediction

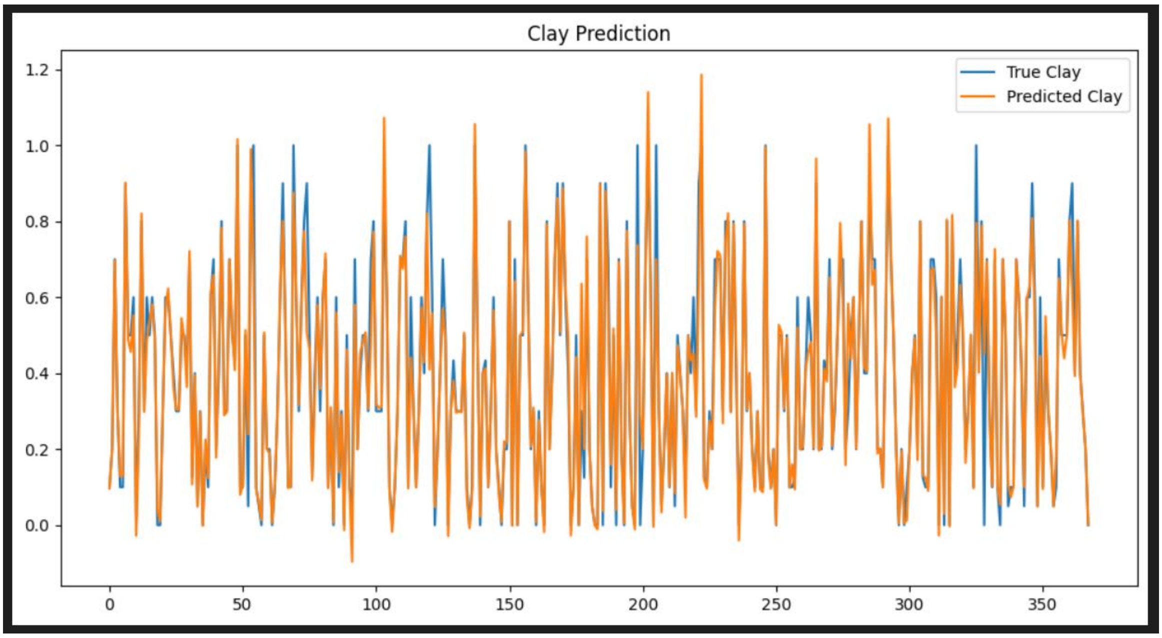


Figure 6 : Clay Prediction

4.3 Soil Classification using the Empirical Triangle

After predicting the proportions of sand, silt, and clay, soil classification is performed using the soil texture triangle—a key tool in soil science. This diagram categorizes soil based on the relative percentages of the three components, with axes representing sand, silt, and clay. By plotting the predicted values within this triangle, the soil is classified into texture types such as clay, loam, or sandy loam, depending on where the values fall. For example, soils with high clay content are categorized in the "Clay" region, while balanced compositions may be classified as "Loam." This classification is essential as it translates raw proportions into meaningful categories that inform water retention, nutrient availability, and crop suitability, ultimately aiding better agricultural decisions.

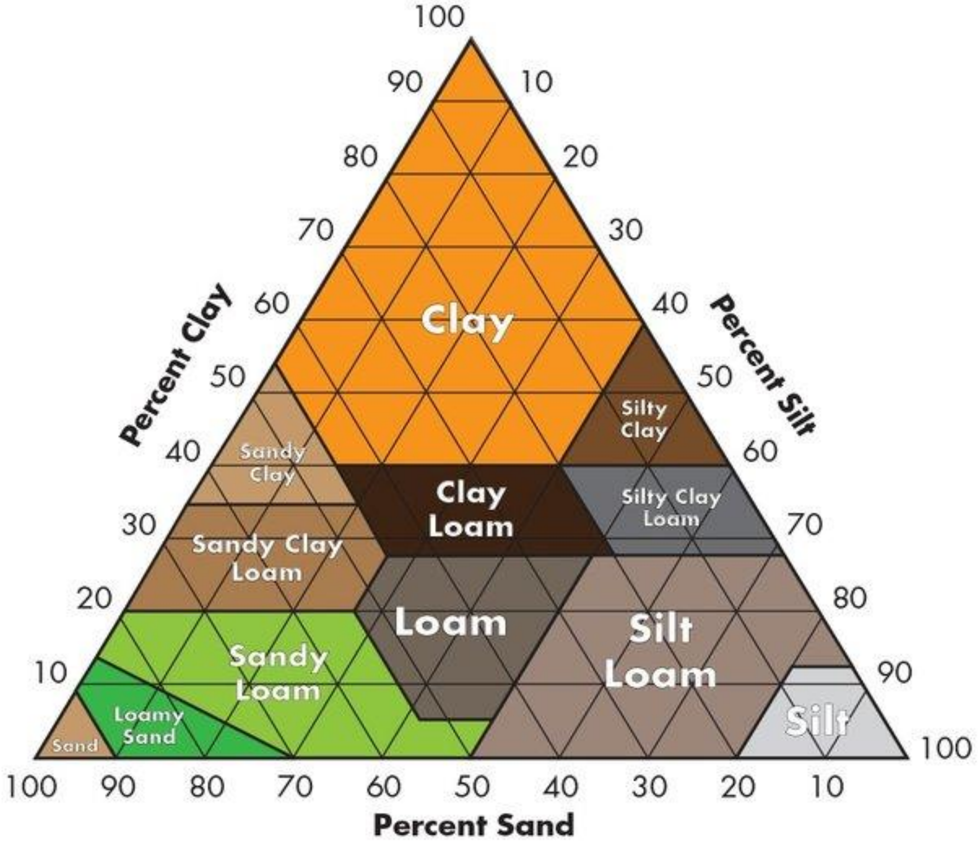


Figure 7 : Soil Triangle

Chapter 5: Conclusion

5.1 Summary of Findings

The Gaussian Process Regression model successfully predicted the proportions of sand, silt, and clay based on multispectral imaging data. The low Mean Squared Error (MSE) value indicates that the model's predictions closely align with actual soil compositions, validating the effectiveness of the approach.

5.2 Contributions and Significance

The results demonstrate that our approach can accurately predict soil composition using multispectral imaging data. This method offers a practical alternative to traditional soil analysis, providing rapid, non-invasive predictions that can be valuable in precision agriculture. Accurate predictions of sand, silt, and clay proportions can inform better crop management practices, leading to improved yield and resource utilization. By integrating advanced image processing, feature extraction, and regression techniques, this study provides a robust framework for soil analysis, with potential applications extending across various agricultural and environmental monitoring tasks.

5.3 Limitations of the Study

While the model performed well overall, some limitations include the presence of outliers and variability within certain soil types that were less accurately predicted. Additionally, the model's performance could be further enhanced by incorporating more diverse datasets representing a wider range of soil conditions.

5.4 Future Work and Recommendations

Future research could explore more advanced modeling techniques, expand the dataset to include more diverse soil types, and refine the image preprocessing pipeline to address any inconsistencies observed. Additionally, integrating other data sources such as environmental factors or satellite imagery could further improve prediction accuracy.

References

- Gopi, E. S., & Palanisamy, P. (2013). *Fast computation of PCA bases of image subspace using its inner-product subspace*. *Applied Mathematics and Computation*, 219(12), 6729-6732. https://doi.org/10.1016/j.amc.2013.01.060

- Gopi, E. S. (2020). *Pattern Recognition and Computational Intelligence Techniques Using Matlab*. [Springer].

- Kempen, B., Brus, D. J., & Heuvelink, G. B. (2009). "Soil Texture Mapping Using Gaussian Process Regression." *Geoderma*, 151(3-4), 243-251.

- Xie, X., & Zhang, J. (2010). "Image Segmentation Using Fast Principal Component Analysis." *Pattern Recognition Letters*, 31(5), 494-501.

- Sun, Z., Song, Y., & Xu, X. (2015). "Remote Sensing Data Processing Using PCA for Soil Property Prediction." *Environmental Modelling & Software*, 67, 128-136.

- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.