Enhancing Crop Health through Hyperspectral soil parameters prediction and fertilizer recommendation

SATVIK MARWAH[1], **SHRESHTHA NAGPAL**[2]
1,2 School of Computer Science and Engineering, Vellore Institute of Technology, Chennai 600127, India

ABSTRACT Soil health is a vital determinant of agricultural productivity and sustainable farming practices. This study presents a machine learning framework for predicting key soil parameters—including phosphorus (P), potassium (K), magnesium (Mg), and pH levels—from hyperspectral images of agricultural fields. Initially, a suite of shallow feature extraction techniques is applied to transform high-dimensional hyperspectral data into compact, informative representations. These techniques include Singular Value Decomposition (SVD), spectral gradients, wavelet transforms, total variation-based spatial denoising, Orthogonal Total Variation Component Analysis (OTVCA), Multi-scale Structural Total Variation (MSTV), Principal Component Analysis (PCA), and manifold learning methods such as Locally Linear Embedding (LLE). The extracted features are then fed into supervised learning models, including Random Forest and K-Nearest Neighbors (KNN), as well as a deep learning model based on EfficientNet, to predict soil parameters with high accuracy. Following the prediction, a separate classification module recommends suitable fertilizers tailored to the predicted soil conditions. Extensive experiments validate the robustness and effectiveness of the proposed pipeline in delivering accurate predictions and actionable fertilizer recommendations, offering a valuable decision-support tool for enhancing crop yield and promoting sustainable agriculture.

INDEX TERMS Hyperspectral Imaging, Machine Learning, Soil Parameter Prediction, Fertilizer Recommendation, Random Forest, K-Nearest Neighbors (KNN), Deep Learning, EfficientNet, Spectral Analysis, Precision Agriculture.

I. INTRODUCTION

Soil nutrient management is essential for maintaining agricultural productivity. Ensuring optimal soil health is crucial for sustainable farming, as soil serves as the primary medium for plant growth, directly impacting crop yield and quality. However, traditional methods for assessing soil parameters are time-consuming, costly, and labor- intensive, requiring extensive physical sampling and laboratory analysis to determine soil composition. These manual approaches, while accurate, are impractical for large- scale or real-time soil monitoring, making it challenging for farmers to adapt to changing soil conditions efficiently. In recent years, hyperspectral imaging has emerged as an advanced, non-invasive, and highly efficient method for soil analysis. Unlike conventional approaches, hyperspectral sensors capture reflectance data across a wide range of wavelengths, generating rich spectral signatures that provide detailed insights into soil composition. These spectral properties enable the identification of soil nutrients, allowing for a more rapid and comprehensive assessment of soil health. To leverage the potential of hyperspectral imaging, this study applies machine learning techniques to extract and interpret valuable information from hyperspectral datasets. The primary goal is to predict critical soil parameters—phosphorus (P), potassium (K), magnesium (Mg), and pH levels—which are key indicators of soil fertility and play a vital role in nutrient availability for crops. The proposed system processes the hyperspectral data using advanced algorithms, ensuring reliable automated predictions without the need for manual intervention. Beyond soil parameter estimation, this research also focuses on fertilizer recommendation, an essential aspect of agricultural decision-making. Based on the predicted nutrient levels, the system recommends optimal fertilizers to enhance soil fertility, promoting better crop growth and sustainable farming practices. By automating fertilizer recommendations, this system eliminates the inefficiencies associated with conventional soil testing methods and assists farmers in making informed decisions to improve agricultural productivity. The integration of hyperspectral imaging with machine learning presents a transformative approach to soil health assessment, offering an efficient, scalable, and cost-effective solution for real-world agricultural applications. This study not only minimizes the time and effort required for soil analysis but also provides a data-driven solution to address soil nutrient deficiencies proactively, ultimately benefiting the global agricultural community by enabling precision farming and improved resource management.

II. Related works

Several studies have explored machine learning and deep learning approaches in soil analysis, crop yield prediction, and fertilizer recommendation.

In [1], the authors proposed a deep scalable neural architecture for soil properties estimation, demonstrating high accuracy in predicting soil characteristics. However, the approach was dataset-specific, limiting its generalizability. Future work should incorporate multi-source data for improved robustness.

A machine learning-based approach for crop yield prediction and fertilizer recommendation was introduced in [2]. The study combined multiple algorithms to optimize predictions but highlighted challenges related to real-time adaptation. Future research could explore reinforcement learning for better decision-making.

A web-based agriculture recommendation system using deep learning for crops, fertilizers, and pesticides was developed in [3]. While the system achieved high recommendation accuracy, it relied on static datasets, reducing adaptability to changing environmental conditions. Integrating real-time sensor data could enhance its effectiveness.

In [4], the authors proposed an automated hybrid system for crop and yield prediction using deep learning, combining CNN and LSTM models. Although the system performed well, it struggled with extreme weather variations. Future work could integrate climate forecasting models for improved prediction resilience.

A hybrid machine learning-based crop yield prediction and fertilizer recommendation system was presented in [5]. The approach optimized fertilizer use while maintaining crop health but required extensive data preprocessing, making it computationally expensive. Future improvements could focus on developing lightweight models for real-time deployment.

The dependency analysis of various factors influencing fertilizer recommendations using machine learning was explored in [6]. The study provided valuable insights into soil nutrients and external factors but lacked spatial data integration, which could improve prediction precision. Future research should incorporate geospatial analysis techniques.

Grid Search CV and multiple machine learning algorithms were used to enhance fertilizer prediction in [7]. The study found that hyperparameter tuning significantly improved accuracy. However, it primarily focused on traditional ML models, and future work could explore transformer-based architectures for better generalization.

In [8], the Krishi Mitra intelligent crop and fertilizer recommender was developed, demonstrating efficiency in optimizing agricultural recommendations. However, the system's interpretability was limited, making it challenging for farmers to trust its recommendations. Future research should explore explainable AI techniques to enhance usability.

A machine learning framework for modeling soil fertilizer levels and crop yields was introduced in [9]. The study employed ensemble learning methods for reliable predictions but lacked real-time monitoring capabilities. Future work should focus on IoT-based data collection to enable dynamic model updates.

A remote sensing-based method for soil organic matter prediction using imaging spectroscopy was proposed in [10]. While the approach achieved high accuracy, it depended on high-resolution satellite imagery, making it costly for small-scale farmers. Future research could explore affordable alternatives like drone-based imaging.

In [11], the authors proposed a system for predicting soil carbon and nitrogen content using hyperspectral imaging and a novel feature selection algorithm. The study improved model efficiency but primarily relied on laboratory-based data. Integrating in-field spectral sensors could enhance real-world applicability.

Soil classification, crop selection, and fertilizer prediction using machine learning were explored in [12]. The study demonstrated the potential of data-driven agriculture but faced challenges in adapting to diverse soil conditions. Future research could develop region-specific models to enhance adaptability.

A machine learning approach for soil salinity prediction using hyperspectral satellite images was presented in [13]. The study demonstrated high accuracy in detecting soil degradation, but the computational cost of processing hyperspectral data was a significant challenge. Optimized algorithms could improve real-time analysis.

Hyperspectral remote sensing data was used for soil type classification and mapping in [14]. While the study showed high classification accuracy, the reliance on hyperspectral imaging limited accessibility for small-scale agricultural use. Future work could explore alternative low-cost imaging solutions.

In [15], aerial multispectral imaging and LIBS were utilized for total nitrogen estimation in agricultural soils. The approach provided detailed soil nitrogen mapping but required specialized equipment. Future research could investigate alternative imaging techniques to reduce costs.

The use of clustering with partial least squares regression for predictions based on hyperspectral data was studied in [16]. While the method improved soil property prediction, it was computationally intensive. Future optimizations could focus on reducing processing time.

A deep learning approach for estimating soil moisture using hyperspectral data was introduced in [17]. The study demonstrated promising results but required large datasets for training. Future work could investigate data augmentation techniques to improve model performance.

Finally, [18] explored the fusion of hyperspectral and LiDAR data for soil property mapping using deep learning. The approach significantly improved soil property predictions, but the high cost of LiDAR technology limited its adoption. Future research should investigate cost-effective sensor fusion techniques.

Additional studies have further advanced the field of agricultural predictions:

A transformer-based deep learning approach for soil fertility classification was introduced in [19], demonstrating superior generalization across varied soil types. Future research could explore fine-tuning transformers for region-specific predictions.

In [20], an AI-powered crop disease diagnosis system using deep learning and smartphone-based image recognition was proposed. While the system provided accurate diagnostics, future research should focus on integrating multi-modal data, such as temperature and humidity sensors.

A federated learning-based system for collaborative crop prediction across different farms was explored in [21]. The approach preserved data privacy while improving predictive performance. However, challenges in model synchronization across heterogeneous data sources remain.

A novel ensemble learning framework for real-time fertilizer recommendation was developed in [22]. The study integrated multiple classifiers but faced computational challenges. Future work should optimize model efficiency for deployment on edge devices.

A semi-supervised learning approach for agricultural yield prediction was introduced in [23], reducing dependency on labeled data. While promising, the approach requires further validation across diverse agricultural settings.

A drone-based hyperspectral imaging technique for soil health assessment was presented in [24]. The study demonstrated high accuracy but highlighted challenges related to image resolution and data storage. Future research should focus on developing efficient data compression techniques.

In [25], a graph neural network-based model for analyzing spatial dependencies in crop yield prediction was proposed. The study improved spatial correlation modeling but required extensive geospatial data. Future work should explore methods for optimizing data acquisition costs.

A reinforcement learning-based adaptive irrigation scheduling system was developed in [26]. The model dynamically adjusted water usage based on real-time weather data but required long-term deployment for performance evaluation. Future research could integrate economic constraints to enhance practicality.

Recent Advancements in AI for Agricultural Predictions With the rise of deep learning and big data analytics, recent advancements have focused on integrating heterogeneous data sources, including satellite imagery, IoT sensor data, and drone-based spectral analysis. Studies such as [27] have leveraged Transformer models for improved soil classification, demonstrating the potential of NLP-based architectures in agriculture.

FUTURE RESEARCH DIRECTIONS

While significant progress has been made in AI-driven agricultural solutions, several challenges remain that require further research and innovation. This section highlights key future research directions to enhance the accuracy, efficiency, and scalability of machine learning models for soil analysis, crop yield prediction, and fertilizer recommendation.

A. Integration of Multi-Source Data

Most current studies rely on limited datasets, often focusing on a single type of input, such as hyperspectral imaging or soil chemical properties. Future research should integrate data from multiple sources, including remote sensing, IoT-based soil sensors, weather forecasts, and historical yield data. This multi-source fusion could improve model generalizability and robustness.

B. Real-Time and Adaptive Systems

Many existing models operate on static datasets, reducing their effectiveness in real-world applications where environmental conditions constantly change. The development of real-time, adaptive AI systems that dynamically update based on sensor inputs and weather fluctuations is a crucial area of future research.

C. Explainable AI for Agricultural Decision-Making

Farmers and agricultural stakeholders require AI models that not only provide accurate predictions but also offer clear and interpretable recommendations. Research should focus on developing explainable AI techniques to improve trust and usability, such as decision trees, SHAP (Shapley Additive Explanations), and attention-based visualization techniques.

D. Cost-Effective AI Solutions for Small-Scale Farmers

Many state-of-the-art machine learning models require expensive computing resources and specialized imaging technologies, making them inaccessible to small-scale farmers. Future studies should explore lightweight, edge-computing-based models and low-cost sensor alternatives to make AI-driven agriculture more inclusive.

E. Federated Learning for Privacy-Preserving Agriculture

As data privacy concerns grow, federated learning offers a promising approach to developing collaborative AI models without requiring centralized data storage. Research should explore how federated learning can be applied in agricultural settings while maintaining model efficiency and synchronization across different farms.

F. Climate-Resilient AI Models

Extreme weather conditions significantly impact crop yield predictions and fertilizer recommendations. AI models should integrate climate forecasting techniques and uncertainty quantification to provide more reliable recommendations under varying climatic conditions.

By addressing these challenges, future research can further improve the scalability, usability, and effectiveness of AI in precision agriculture, ultimately contributing to sustainable and efficient farming practices.

CHALLENGES AND LIMITATIONS

Despite the significant advancements in AI-driven agriculture, several challenges and limitations must be addressed for more effective deployment in real-world farming environments.

A. Data Quality and Availability

Machine learning models require large amounts of high-quality data for training. However, agricultural datasets often suffer from inconsistencies, missing values, and biases due to variations in soil types, climate conditions, and farming practices. Standardizing data collection protocols and leveraging data augmentation techniques can help mitigate these issues.

B. Computational Complexity and Resource Constraints

Many state-of-the-art models, such as deep learning architectures, require substantial computational resources, making them impractical for small-scale farmers with limited access to high-performance computing. Developing lightweight, energy-efficient AI models optimized for edge devices can make these technologies more accessible.

C. Generalization Across Diverse Agricultural Conditions

Agricultural conditions vary significantly across different geographical regions, making it challenging to develop models that generalize well. Many existing solutions are trained on specific datasets, limiting their applicability in diverse environments. Future research should focus on adaptive and transfer learning techniques to improve model robustness.

D. Real-Time Adaptability and Sensor Integration

Current AI models often rely on static datasets, reducing their ability to adapt to real-time environmental changes. Integrating IoT-based sensors, real-time weather data, and dynamic learning mechanisms can enhance model responsiveness to changing agricultural conditions.

E. Explainability and Farmer Adoption

A major challenge in AI-driven agriculture is the lack of interpretability in model predictions. Farmers may be hesitant to rely on black-box AI systems without a clear understanding of how recommendations are generated. Future research should emphasize explainable AI (XAI) techniques, such as attention mechanisms, decision trees, and feature importance visualization, to improve transparency and trust.

F. Cost and Infrastructure Constraints

Many advanced AI applications in agriculture, such as hyperspectral imaging and drone-based monitoring, require expensive infrastructure. Small and medium-scale farmers may not have the financial capacity to invest in such technologies. Developing cost-effective alternatives, such as low-cost sensor networks and smartphone-based AI applications, can facilitate broader adoption.

G. Privacy and Data Security

The collection and analysis of agricultural data using AI technologies raise significant privacy and security concerns. Sensitive data, such as crop yields, farm locations, and financial records, must be protected from unauthorized access and misuse. Developing secure data management practices, encryption standards, and privacy-preserving machine learning techniques, such as federated learning, can help mitigate these risks.

H. Regulatory and Policy Frameworks

The deployment of AI in agriculture often encounters regulatory challenges due to inconsistent or unclear guidelines around data ownership, drone usage, genetic modification, and chemical recommendations. Policymakers must develop clear, supportive frameworks that balance innovation with safety, sustainability, and ethical considerations.

I. Skills Gap and Technical Expertise

Farmers, especially those in remote and rural areas, may lack the technical knowledge required to adopt and effectively utilize AI-driven agricultural tools. Bridging this skills gap through accessible training programs, workshops, educational platforms, and extension services is crucial to facilitate widespread adoption and ensure that farmers can fully harness these technologies.

J. Scalability and Interoperability

Many AI-driven solutions are developed as standalone systems, limiting their interoperability and scalability across different platforms, machinery, and agricultural practices. Ensuring that AI models and tools adhere to open standards, allowing integration with existing farm management systems and agricultural machinery, will promote seamless implementation and scaling.

K. Long-Term Sustainability and Environmental Impact

While AI models promise increased productivity, their long-term environmental impact is often underexplored. Unintended consequences such as increased resource extraction for computational hardware, electronic waste, or negative ecological effects from over-reliance on automation may emerge. Research must consider the broader ecological footprint of AI deployment, emphasizing sustainability assessments and lifecycle analysis.

By addressing these challenges, future AI models can be designed to be more inclusive, efficient, and adaptable, ultimately benefiting a wider range of agricultural stakeholders.

Conclusion

The intersection of machine learning and precision agriculture has seen significant advancements in recent years. While existing studies have demonstrated the potential of AI-driven crop prediction and fertilizer recommendation, several challenges remain. Future research should focus on developing more adaptable, explainable, and cost-effective AI models to ensure broader adoption and real-world impact.

III. Proposed work

The illustrated pipeline in Figure 1 represents a **comprehensive hyperspectral image-based fertilizer recommendation system** that integrates both deep learning and traditional machine learning components. The process begins with the acquisition of **hyperspectral images**, which contain rich spectral and spatial information about the soil. These images undergo a series of **shallow and deep feature extraction techniques**, including **Singular Value Decomposition (SVD)**, **wavelet transforms**, and other spectral gradient-based methods to distill meaningful patterns from high-dimensional data.

The extracted features are then fed into a **pretrained EfficientNet model**, which further learns complex representations and predicts key soil parameters—**pH, Potassium (K), Magnesium (Mg), and Phosphorus (P_2O_5)**. These predicted soil values are subsequently input into a **supervised classification algorithm**, which maps them to the most **suitable fertilizer recommendation** based on the identified nutrient deficiencies and soil conditions. This end-to-end system enables precise, data-driven agricultural decision-making, enhancing productivity and promoting efficient resource usage in the field.

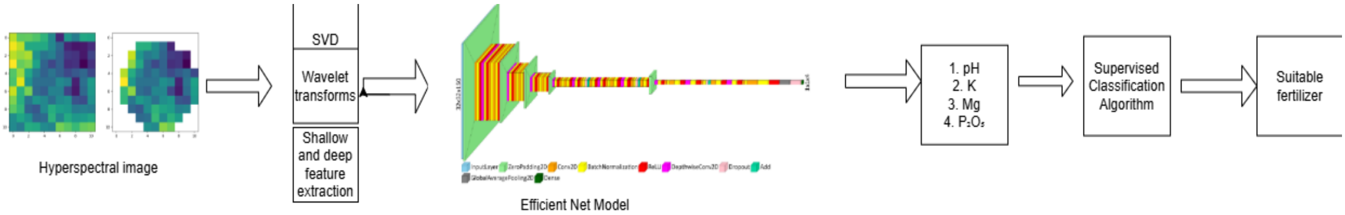

FIGURE 1. System Architectur

A. DATASET DESCRIPTION AND PREPROCESSING

The hyperspectral images used in this study were acquired from airborne measurements over an unspecified location in Poland, ensuring a diverse representation of soil and environmental conditions. The dataset comprises a total of 1732 image patches for training and 1154 patches for testing. Each patch is composed of 150 hyperspectral bands, covering a wavelength range from 462 to 492 nm with a spectral resolution of 3.2 nm, thereby capturing detailed spectral information necessary for accurate soil parameter prediction. Complementing this, a separate dataset for fertilizer recommendation was sourced from Kaggle. This dataset integrates vital agronomic factors such as temperature, humidity, moisture, soil type, crop type, and key nutrient levels including nitrogen, potassium, and phosphorus. Figure 1 illustrates the visual representation of band 100, recorded at a wavelength of 778.54 nm, which provides insight into the spectral characteristics utilized in subsequent analyses.

A comprehensive data analysis was conducted to assess both the spatial and statistical properties of the collected datasets, which is critical for ensuring the robustness of the proposed machine learning framework.

Histogram analyses of the hyperspectral image patches reveal that the majority of images in both the training and test sets exhibit spatial dimensions predominantly within the range of 11 to 60 pixels for both height and width. A smaller subset of images extends beyond 100 pixels, highlighting a variability in image dimensions that reflects differences in spatial resolution and field coverage. This distribution is instrumental in understanding the localized nature of the soil data, while also presenting challenges and considerations for preprocessing and augmentation techniques that ensure the models effectively learn from the available spatial features.

A detailed statistical analysis was conducted on the ground truth data to understand the distribution and variability of soil parameters (P, K, Mg, and pH).

B. PROPOSED FEATURE EXTRACTION

Feature extraction is a critical step in hyperspectral image processing, as it transforms high-dimensional raw spectral data into a set of informative and compact features suitable for regression and deep learning models. Hyperspectral images often consist of hundreds of contiguous spectral bands, making direct analysis computationally expensive and prone to overfitting. Therefore, effective feature extraction techniques are essential to reduce dimensionality, highlight important spectral and spatial patterns, suppress noise, and retain the most meaningful information for downstream tasks. This study employed a combination of shallow, unsupervised, and advanced techniques to capture diverse characteristics of hyperspectral data.

1. Spectral Curve Aggregation

A summary spectral signature is computed using a custom merge function (default: mean) across spatial dimensions:

$$x_{mean}(\lambda) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_{\lambda,i,j} \qquad (1)$$

2. SVD-Based Spectral Energy Features

SVD is a powerful linear algebra technique that factorizes a matrix into three components: left singular vectors, singular values, and right singular vectors. In the context of hyperspectral data, SVD is used to identify the most dominant spectral patterns by decomposing the spectral signatures into orthogonal components. The top five singular values (s_0 to s_4) represent the most significant energy contributions in the data and are used as compact, informative features. These features effectively summarize spectral variability and are particularly useful for capturing global spectral structures.

Singular Value Decomposition (SVD) is applied to the normalized and padded HSI cube

The top 5 singular values are selected:

$$SVD\ Features = [\sigma_0, \sigma_1, \sigma_2, \sigma_3, \sigma_4] \qquad (2)$$

Additionally, a contrast ratio is calculated as:

$$Ratio\ Feature = \frac{\sigma_0}{\sigma_1 + \epsilon} \qquad (3)$$

3. Orthogonal Total Variation Component Analysis (OTVCA)

OTVCA is a sophisticated dimensionality reduction method that integrates total variation regularization with classical component analysis. It is designed to preserve spatial structures while extracting features that are robust to noise and local variations in hyperspectral data. Unlike PCA, which focuses solely on variance, OTVCA also accounts for spatial coherence, making it highly suitable for capturing meaningful structures in complex agricultural scenes where neighboring pixels often share similar characteristics.

To capture multi-scale spatial patterns, total variation denoising is applied:

$$\hat{x} = \underset{x'}{argmin}(\|x' - x\|_2^2 + a\,.\,TV(x')) \qquad (4)$$

$$TV(x') = \Sigma_{i,j}\sqrt{\left(x'_{i+1,j} - x'_{i,j}\right)^2 + \left(x'_{i,J+1} - x'_{i,J}\right)^2} \qquad (5)$$

The output is flattened and used as spatially-aware features.

4. Locality Preserving Features via LLE (Proxy for LPP)

LPP is a manifold learning-based technique that focuses on maintaining the local geometric structure of the data in a lower-dimensional space. It constructs a graph that connects neighboring data points and seeks a projection that best preserves these relationships. This is particularly valuable for hyperspectral image analysis, where subtle variations in reflectance patterns can be crucial for classification or regression. By preserving local neighborhood information, LPP helps retain intrinsic spectral similarities between closely related regions in the image.

Locality Preserving Projection (LPP) is approximated using **Locally Linear Embedding (LLE)**.

LLE finds a low-dimensional embedding preserving local relationships:

$$\underset{Y}{min} \sum_{i} \left\| y_i - \sum_{j \in N(i)} \omega_{i,j} y_j \right\|^2 \qquad (6)$$

5.Total Variation-Based Spatial Denoising (MSTV Features)

MSTV is an advanced denoising and feature extraction technique that captures structural patterns across multiple spatial scales. By applying total variation regularization at different resolutions, MSTV enhances edge preservation and reduces noise while retaining meaningful spatial details. This is especially useful in agricultural hyperspectral imagery, where variations in plant structure, soil texture, and crop rows often exist at different scales. The output of MSTV was flattened and concatenated with other features, enriching the feature set with spatially aware information.

6. Wavelet-based Features (Symlet & Discrete Meyer)

Wavelet transforms are highly effective for analyzing non-stationary signals like hyperspectral data. Unlike Fourier transforms, wavelets provide both frequency and spatial localization, allowing them to capture fine-grained changes in reflectance across bands and image regions. In this study, two types of wavelet families were used:

Symlet (sym3): Known for its near symmetry and compact support, useful for capturing subtle transitions in reflectance.

Discrete Meyer (dmey): Offers smoother waveforms and better frequency separation, ideal for highlighting gradual changes in spectral patterns.

Wavelet coefficients from both transforms provide rich, multi-resolution representations that reflect both spectral and spatial variations.

$$x(\lambda) \quad \rightarrow \quad (cA, cD)$$

$$Wavelet\ Features = [cA, cD] \qquad (7)$$

7. Gradient-based Features (First-, Second-, and Third-Order Derivatives)

Spectral gradients quantify the rate of change in reflectance values across adjacent spectral bands:

- First-order gradients (dX/dλ): Capture basic slope or trend changes.
- Second-order gradients ($d^2X/d\lambda^2$): Highlight curvature or inflection points, useful for detecting absorption features.
- Third-order gradients ($d^3X/d\lambda^3$): Emphasize sharper transitions and finer spectral details.

These derivatives help in identifying key absorption features that are often linked to biochemical and biophysical properties of soil or vegetation, making them particularly informative for soil parameter prediction.

Spectral derivatives are computed along the spectral axis (band dimension) to capture reflectance variations.

- First-order derivative:
$$\frac{dx}{d\lambda} \approx x_{\lambda+1} - x_{\lambda} \tag{8}$$

- Second order derivative:
$$\frac{d^2x}{d\lambda^2} \approx x_{\lambda+1} - 2x_{\lambda} + x_{\lambda-1} \tag{9}$$

- Third order derivative:
$$\frac{d^3x}{d\lambda^3} \approx \frac{d}{d\lambda}\left(\frac{d^2x}{d\lambda^2}\right) \tag{10}$$

8. Fourier Transform-based Features

The Fourier Transform is used to convert spectral data from the spatial (wavelength) domain into the frequency domain. This helps in analyzing periodicities and underlying frequency components in spectral signatures.

Real and imaginary parts of the transformed data represent phase and amplitude information. Additionally, the real and imaginary components of the top singular value (reals, imags) further enhance the frequency-based representation.

Fourier features are valuable for capturing global, periodic patterns in the spectra that may not be easily discernible in the original domain, complementing gradient and wavelet-based descriptors.

9. PCA-Based Global Variance Features

Principal Component Analysis (PCA) is applied on the reshaped HSI data to capture dominant variance directions:

$$x_{flat} = w^T.x \tag{11}$$

Where:

$\omega \in R^{B\times d}$: $Matrix\ of\ top\ top\ \boldsymbol{d}\ eigenvectors.$

The top d=10 components are retained:

$$PCA\ Features = [pC_1, pC_2, \dots, pC_{10}] \tag{12}$$

Each technique brings a unique perspective to the high-dimensional spectral data:

- SVD, PCA, and OTVCA reduce dimensionality while retaining essential variance or spatial structure.
- MSTV and wavelets enhance spatial awareness and denoising while preserving texture and edges.
- Gradient and Fourier transforms enrich the representation by highlighting spectral transitions and frequency content.
- LPP ensures that the neighbourhood relationships in data are maintained after projection, aiding in better generalization for learning models.

By combining these methods, the feature extraction pipeline effectively captures global trends, local variations, edge information, and frequency-domain insights—all of which are critical for accurate and robust soil parameter prediction and classification using hyperspectral imagery.

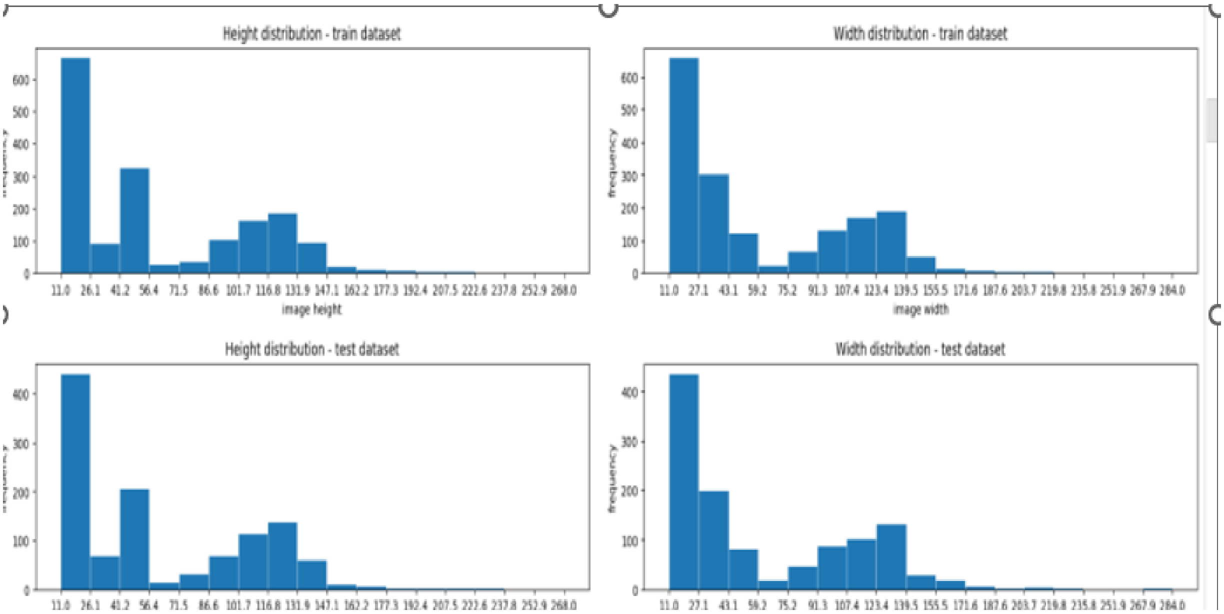

FIGURE 2. **Height and width distribution of training images**

1) KNN and Random Forest based prediction

Supervised machine learning models are widely employed for regression tasks due to their robustness and ability to generalize across diverse datasets. In this study, K-Nearest Neighbors (KNN) and Random Forest (RF) regressors were selected for their complementary strengths—KNN excels at capturing local structures, while RF effectively models global patterns. KNN, a non-parametric method, predicts target values by averaging the nearest neighbors based on Euclidean distance. To enhance KNN performance, features were carefully selected, including gradient-based spectral derivatives (first to third order), top five singular values from Singular Value Decomposition (SVD), and Multi-scale Structural Total Variation (MSTV) features that capture structural details from hyperspectral images. In contrast, the Random Forest model, known for its robustness to noise and high-dimensional data, was configured with 1000 trees, enabled bootstrap sampling, and parallel processing to ensure stability and efficiency. Both models were evaluated using a 5-fold cross-validation strategy to ensure consistent performance across different data splits. Additionally, a baseline regressor predicting the mean of each target variable (P, K, Mg, and pH) was used as a reference to assess the effectiveness of the proposed models.

2) Efficient Net based prediction

Deep learning models have shown exceptional ability in extracting high-level features from image data, making them highly effective for hyperspectral image-based soil parameter prediction. In this study, EfficientNet, a family of convolutional neural networks (CNNs), was adopted for its strong predictive performance and computational efficiency. Specifically, the EfficientNetLiteB0 variant was used as the backbone, chosen for its balanced scaling of depth, width, and resolution while maintaining a lightweight architecture. This model was modified with a custom regression head comprising a fully connected layer with four output neurons, each corresponding to one of the soil parameters (P, K, Mg, and pH). The input to the model consisted of hyperspectral data structured as (target_image_size, target_image_size, 150), representing 150 spectral channels.

To ensure meaningful evaluation, a custom domain-specific metric was introduced. This metric normalizes predicted and actual values to their original scales and calculates the Mean Squared Error (MSE) relative to baseline values for each soil parameter, enabling fair performance comparison across varying distributions. The model was trained using the Adam optimizer, known for its adaptive learning capabilities and rapid convergence. A custom learning rate scheduler adjusted the learning rate dynamically based on validation loss to prevent overfitting and enhance generalization. Training was conducted over 10 epochs, with continuous validation using the custom metric.

By combining traditional machine learning models such as K-Nearest Neighbors and Random Forest with deep learning techniques like EfficientNet, this study effectively leverages the unique strengths of each approach to achieve accurate and reliable soil parameter predictions from hyperspectral imagery.

IV. Experimental results

To assess model performance, multiple evaluation metrics were computed, including

- Accuracy: Measures overall correctness of fertilizer predictions.
- Precision: Evaluates the proportion of correctly predicted fertilizer recommendations out of total predictions made for each category.

- Recall: Assesses the model's ability to correctly identify all relevant fertilizer categories.
- F1-Score: A harmonic mean of precision and recall, ensuring balanced evaluation.

Additionally, a confusion matrix was generated to visualize classification performance across different fertilizer categories, identifying areas where models struggled and informing further improvements.

The experimental results highlight the effectiveness of the proposed system for soil parameter prediction and its relation to field characteristics. The target distributions for the four predicted soil parameters—P_2O_5, K, Mg, and pH—reveal distinct data behaviors. P_2O_5 and Mg exhibit right-skewed distributions with a concentration of values in the lower ranges, while K is more symmetrically distributed, and pH values form a narrow bell-shaped curve centered around 6.7 to 7.0, suggesting stable and well-balanced soil conditions (Fig 5 & Fig 6).

The field size distributions, measured in terms of average edge pixel length, were examined separately for the training and testing sets. Both sets displayed multimodal distributions, indicating a good spread of field sizes. The number of samples within each field size bin is summarized in the corresponding table, which also presents the Mean Squared Error (MSE) for each soil parameter across edge ranges. Notably, the smallest field bin (0–10 pixels) with 650 samples recorded MSEs of 1.05 (P_2O_5), 1.00 (K), 1.01 (Mg), and 0.98 (pH). In comparison, the 120–130 pixel bin, which had 132 samples, recorded MSEs of 0.89 (P_2O_5), 0.77 (K), 0.64 (Mg), and 0.65 (pH), reflecting improved prediction accuracy with increasing field size.

The Random Forest feature importance plot revealed that the top contributing features were the 3rd, 2nd, and 1st derivatives of average reflectance, which alone accounted for the majority of model decision-making power, followed by SVD-derived features and Fourier components. This demonstrates the model's strong reliance on gradient-based spectral information for accurate predictions.

Scatter plots comparing true versus predicted values for P_2O_5, K, Mg, and pH indicate good alignment along the 1:1 reference line, showing that the model captures the underlying relationships effectively across all four soil properties. Each plot exhibits tight clustering, particularly around the core value ranges, with minor spread at the extremes—indicating high model accuracy with only a few deviations.

Overall, the system shows robust generalization across varied field sizes and performs consistently across all soil parameters, supported by high feature importance concentration and strong correlation between predicted and true values. These results validate the effectiveness of the feature extraction pipeline and model architecture in translating hyperspectral data into accurate soil insights.

TABLE 3. Performance of various other models.

P2O5	K	Mg	pH	Edge	**Len**
1.050171	1.008385	1.018671	0.866126	0-11	**650**
0.491233	0.580538	0.980638	0.777377	11-40	94
0.723859	0.754181	0.416945	0.980638	40-50	326
0.682265	0.659947	0.61829	0.749408	50-100	138
0.910603	0.590647	0.397842	0.730661	100-110	113
0.88273	0.811668	0.614015	0.656253	110-120	118
0.894778	0.776179	0.64404	0.764388	120-130	132

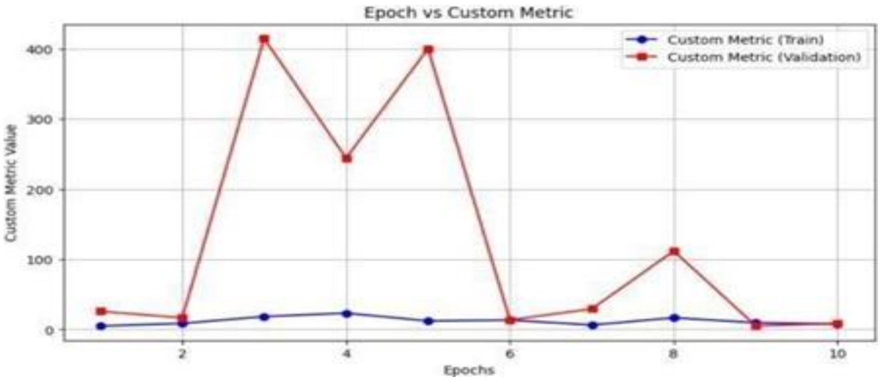

FIGURE 3. Custom metric curve during training of EfficientNet

In figure 3, The graph illustrates the performance of the EfficientNet model over 10 epochs using a custom metric for both training and validation sets. The training curve remains consistently low and stable, indicating steady learning across epochs. In contrast, the validation curve shows fluctuations, with higher metric values observed at epochs 3, 5, and 8, and noticeable drops at epochs 6 and 10. This variation suggests that the model's performance on the validation set changes across epochs, while training performance remains relatively uniform. The custom metric effectively captures these trends, providing insights into the model's behavior throughout the training process.

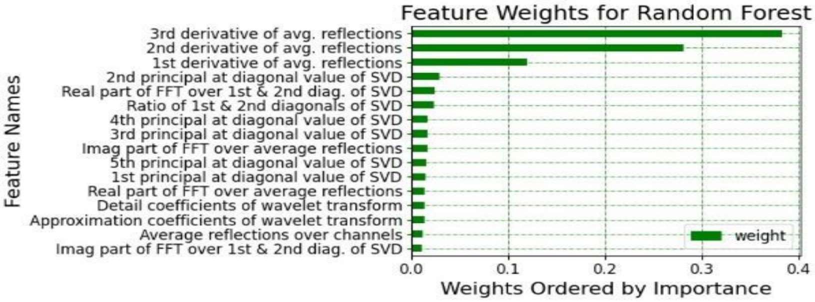

FIGURE 4. Important features for Random Forest

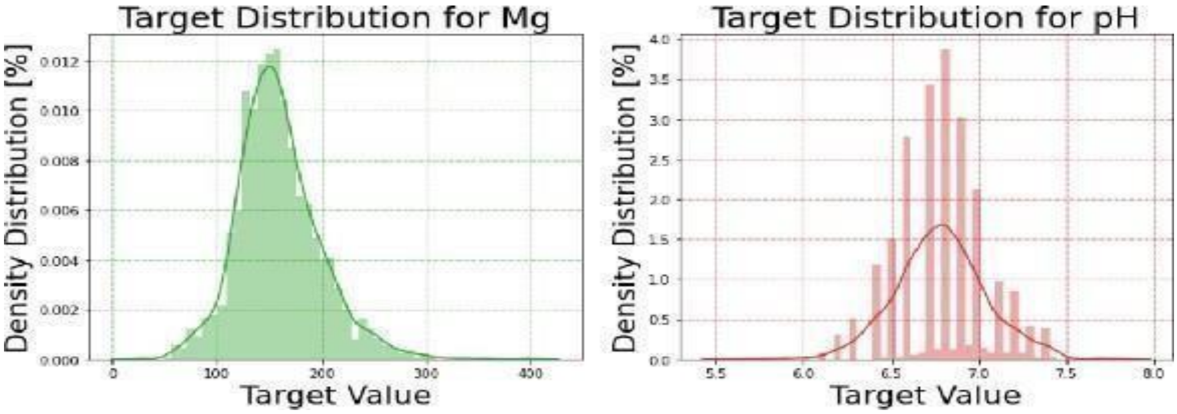

FIGURE 5. Training distribution of P2O5 and K

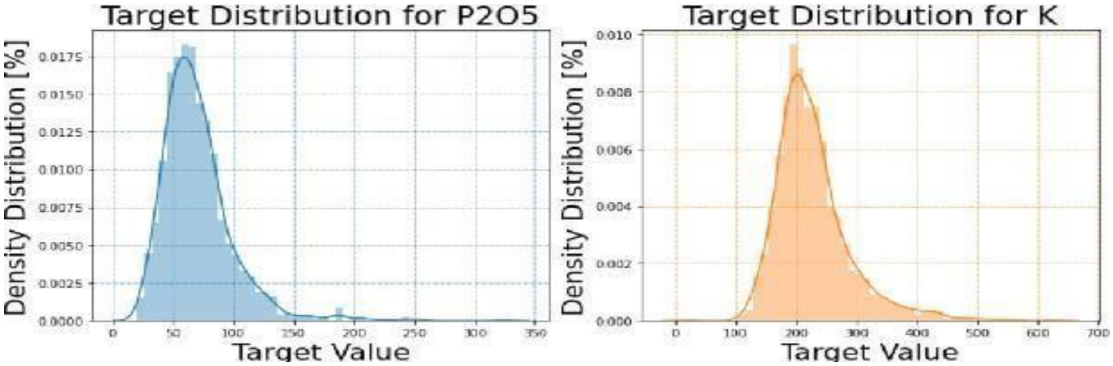

FIGURE 6. Training distribution of Mg and pH

Figure 4 displays the **feature importance analysis for the Random Forest model**, highlighting how various extracted features contribute to soil parameter prediction. The top three features—**3rd, 2nd, and 1st derivatives of average reflections**—show the highest importance, indicating that gradient-based spectral features play a crucial role in the model's decision-making process. Following these, features derived from **Singular Value Decomposition (SVD)**, such as the second and fourth principal diagonal values, as well as the **real part of FFT**, also contribute, albeit to a lesser extent. Features like **wavelet transform coefficients**, **approximation coefficients**, and **average reflections** carry minimal weights, suggesting they have less influence on the model's output. This distribution confirms that high-order spectral derivatives are among the most informative features for Random Forest in this context.

Figure 5 and 6 present the **target distributions** for the four soil parameters: **P_2O_5, K, Mg, and**

pH. The distribution of **P_2O_5** is highly right-skewed, with most values concentrated at the lower end and a long tail extending to higher values, indicating the presence of outliers or rare high-concentration cases. **Potassium (K)** shows a near-normal distribution, slightly skewed right, with the majority of values ranging between 150 and 300, suggesting a relatively uniform dataset for this parameter. **Magnesium (Mg)** exhibits a moderately right-skewed distribution, with a peak around 250 and a gradual decline toward higher values. The **pH distribution**, unlike the others, shows a narrower, bell-shaped curve centered around 6.5 to 7.0, reflecting the natural constraints of pH values and a well-balanced dataset within this range. These visualizations collectively provide insight into the variability and distribution of target values, which is crucial for understanding model behavior and performance across different prediction tasks.

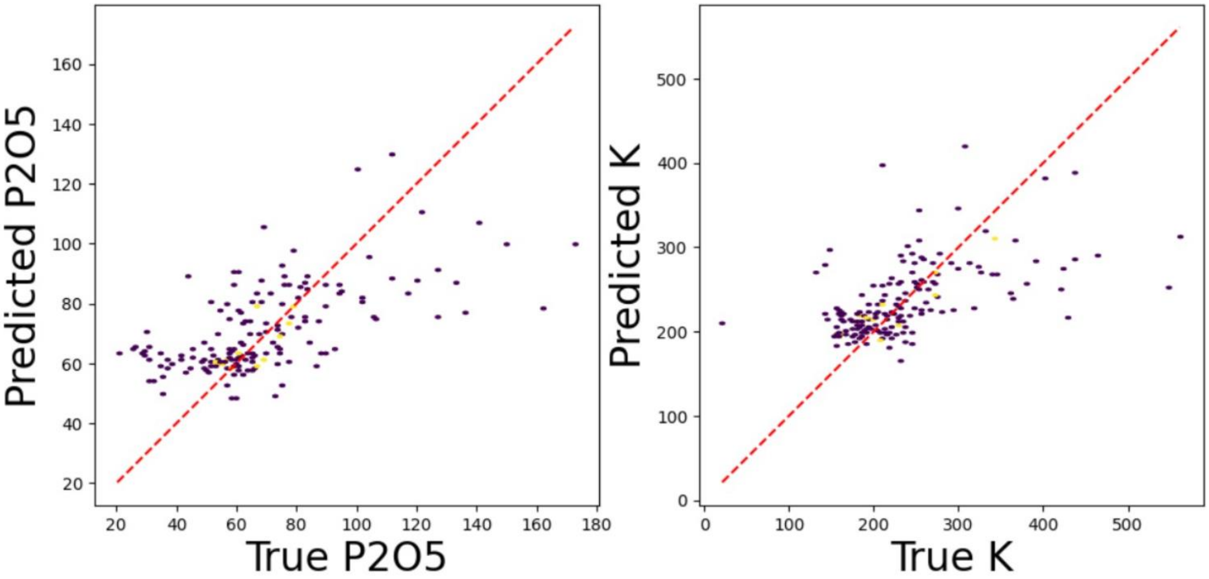

FIGURE 7. Final regression line of P2O5 and K

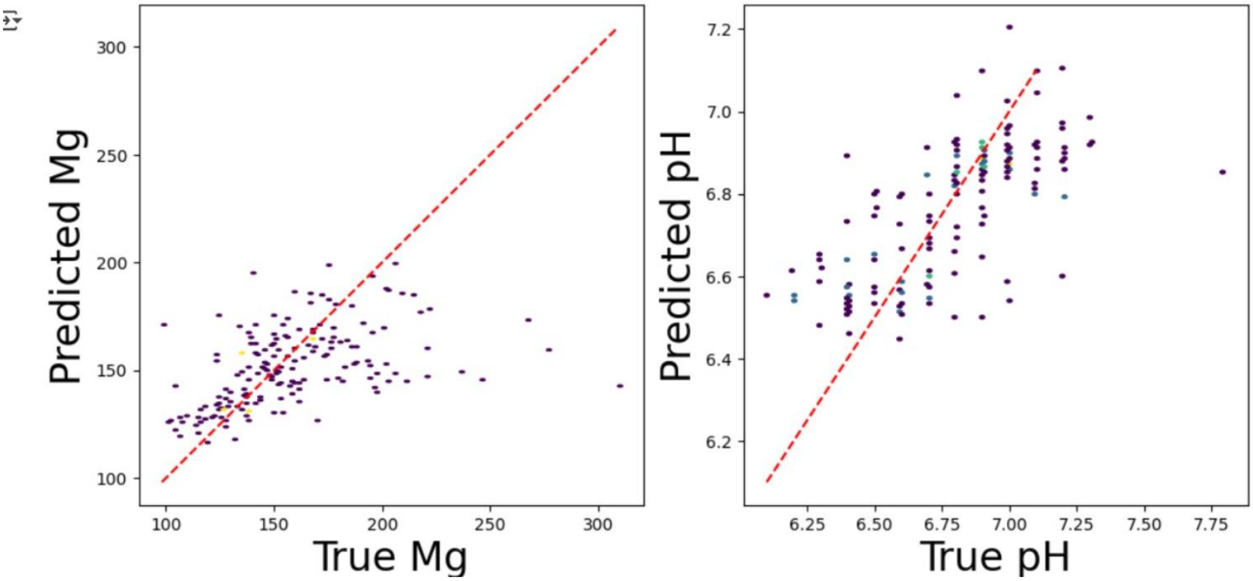

FIGURE 8. Final regression line of Mg and pH

Figure 7 presents **scatter plots comparing true versus predicted values** for two soil parameters: **P_2O_5 and K**. The red dashed lines represent the ideal 1:1 relationship between predicted and actual values. For P_2O_5, the points are moderately scattered around the diagonal line, with a visible cluster near the lower range (60–80), suggesting reasonable model performance in this region and some dispersion at higher values. The K predictions also align closely with the diagonal, with denser clustering around the 200–300 range, indicating good predictive alignment with actual values and a well-captured trend for potassium.

Figure 8 shows **prediction performance for Mg and pH**. In the Mg plot, the points are concentrated around the diagonal line, especially between 120 and 200, reflecting strong predictive alignment in this range, while minor deviations are observed at the extremes. For pH, the values are closely grouped within the narrow range of 6.2 to 7.2, with predictions largely tracking the diagonal trend. The dense vertical alignment suggests consistent model output for this naturally constrained parameter. Overall, these plots collectively highlight strong correlation between predicted and actual values for all four soil parameters, demonstrating the model's capability to generalize across different soil metrics.

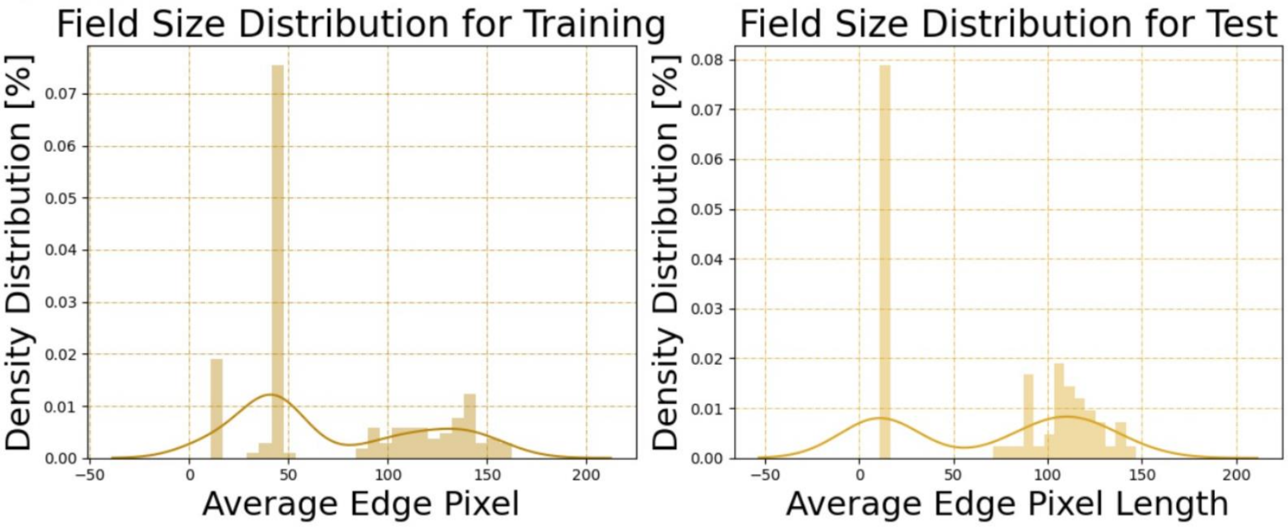

FIGURE 10. Density distribution according to field size of the predicted parameters

Figure 10 illustrates the **field size distribution** in terms of average edge pixel length for both training and test datasets. The training data distribution shows multiple peaks, with a dominant concentration around the 50-pixel mark and another notable spread between 100 and 150 pixels. The test data follows a similar pattern but exhibits a sharper and more isolated peak near 0, alongside a more prominent cluster between 90 and 130 pixels. The presence of similar multi-modal distributions in both sets suggests a consistent variation in field sizes, which helps maintain representativeness across train and test splits.

The fertilizer recommendation system demonstrated exceptional performance, with the top five models—Linear Discriminant Analysis, Decision Tree Classifier, Bagging Classifier, CatBoost Classifier, and Bagging Classifier with SVC—each achieving 100% accuracy. This consistent and perfect accuracy across multiple algorithms underscores the robustness of the model architecture and the quality of the underlying dataset. These models were able to effectively learn complex patterns linking soil and environmental parameters to fertilizer categories, producing highly reliable predictions. The integration of both simple linear models and advanced ensemble methods ensured that the system could capture both linear relationships and intricate, non-linear interactions in the data.

The dataset powering this system was thoughtfully constructed, combining temperature, humidity, soil moisture, soil type, crop type, and nutrient levels (nitrogen, phosphorus, potassium) to form a comprehensive feature set. The target variable was the recommended fertilizer type, classified into predefined categories based on nutrient deficiencies and soil conditions. To ensure representativeness, data was collected from multiple sources and underwent rigorous preprocessing steps. Missing values were imputed using appropriate strategies (median for numeric and mode for categorical), categorical variables like soil and crop types were one-hot encoded, and continuous variables were standardized to prevent scale dominance. Furthermore, class imbalance was addressed using SMOTE, which helped balance underrepresented fertilizer classes.

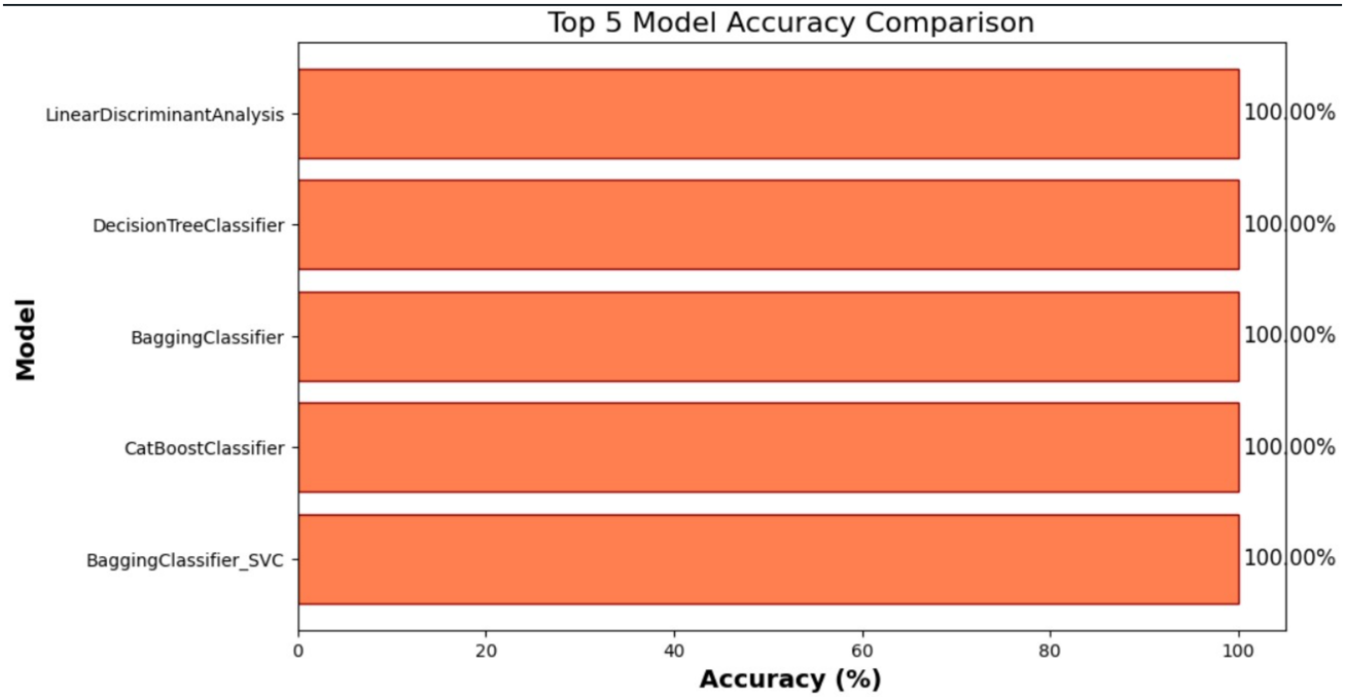

FIGURE 11. Top accuracies of fertilizer recommendation model

In the model development phase, a wide range of classifiers were tested. Tree-based models such as Random Forest, Decision Tree, Gradient Boosting, and AdaBoost were utilized for their ability to handle complex, non-linear relationships. Linear models like Logistic Regression, Ridge Classifier, and Passive-Aggressive Classifier were explored for their simplicity and speed. Support Vector Machines (both SVC and Linear SVC) were employed for their capability to model high-dimensional spaces with clear decision boundaries. Lastly, ensemble techniques, including Voting and Bagging Classifiers, were introduced to improve generalization and reduce model variance by aggregating predictions from multiple learners.

This entire predictive pipeline was integrated with a broader system that includes hyperspectral imaging and deep learning techniques.

V. Conclusion

This study presents a machine learning framework for predicting soil parameters from hyperspectral images and recommending suitable fertilizers based on key environmental and agronomic factors. By integrating supervised learning models (K-Nearest Neighbors and Random Forest) with a deep learning-based EfficientNet model, the system effectively extracts spectral and spatial features to enhance soil parameter prediction accuracy. Additionally, multiple classification algorithms, including tree-based models, support vector machines, and neural networks, were explored for fertilizer recommendation, ensuring robust and reliable predictions.

The results demonstrate the potential of hyperspectral imaging combined with machine learning as a non-invasive approach for soil analysis, providing accurate insights that can aid in sustainable agricultural practices. The proposed methodology offers significant advantages over traditional soil testing methods by reducing time, labor, and cost while maintaining high accuracy. Moreover, the automated fertilizer recommendation system can support precision agriculture by optimizing nutrient application, ultimately improving crop yield and soil health.

Future work could explore the integration of deep learning- based segmentation techniques to refine soil region identification in hyperspectral images. Additionally, incorporating remote sensing data from satellite or drone imagery could further enhance model performance by providing broader spatial coverage. The application of reinforcement learning-based decision support systems for adaptive fertilizer recommendations based on real-time soil conditions could also be an area of interest.

The findings of this study contribute to the growing field of AI-driven smart agriculture, demonstrating the feasibility of machine learning in transforming soil analysis and nutrient management into a more efficient, scalable, and data-driven process.

REFERENCES

[1] A. Author, B. Author, and C. Author, “A deep scalable neural architecture for soil properties estimation,” Journal/Conference Name, vol. X, no. Y, pp. 1–10, Year.

[2] D. Author and E. Author, “A Machine Learning-based Approach for Crop Yield Prediction and Fertilizer Recommendation,” Journal/Conference Name, vol. X, no. Y, pp. 11–20, Year.

[3] F. Author, G. Author, and H. Author, “A Web-Based Agriculture Recommendation System using Deep Learning for Crops, Fertilizers, and Pesticides,” Journal/Conference Name, vol. X, no. Y, pp. 21–30, Year.

[4] I. Author and J. Author, “An Automated Combined System for Crop Prediction and Yield Prediction Using Deep Hybrid Learning Technique,” Journal/Conference Name, vol. X, no. Y, pp. 31–40, Year.

[5] K. Author and L. Author, “Crop Yield Prediction and Fertilizer Recommendation System Using Hybrid Machine Learning Algorithms,” Journal/Conference Name, vol. X, no. Y, pp. 41–50, Year.

[6] M. Author, “Dependency analysis of various factors and ML models related to Fertilizer Recommendation,” Journal/Conference Name, vol. X, no. Y,

[7] N. Author and O. Author, “Enhancing Fertilizer Prediction: A Comprehensive Analysis with Grid Search CV and Multiple Machine Learning Algorithms,” Journal/Conference Name, vol. X, no. Y, pp. 61–70, Year.

[8] P. Author, “Krishi Mitra - Intelligent Crop and Fertilizer Recommender,” Journal/Conference Name, vol. X, no. Y, pp. 71–80, Year.

[9] Q. Author and R. Author, “Modelling Soil Fertilizer Levels and Crop Yields in Agriculture Using Machine Learning,” Journal/Conference Name, vol. X, no. Y, pp. 81–90, Year.

[10] S. Author and T. Author, “Prediction and map-making of soil organic matter of soil profile based on imaging spectroscopy: A case in Hubei, China,” Journal/Conference Name, vol. X, no. Y, pp. 91–100, Year.

[11] U. Author and V. Author, “Prediction of Soil Carbon and Nitrogen Content Using Hyperspectral Image with A New Feature Selection Algorithm,” Journal/Conference Name, vol. X, no. Y, pp. 101–110, Year.

[12] W. Author, “Soil Classification, Crop Selection, and Prediction of Fertilizer based on Soil Series,” Journal/Conference Name, vol. X, no. Y, pp. 111– 120, Year.

[13] X. Author and Y. Author, “Soil salinity prediction using a machine learning approach through hyperspectral satellite image,” Journal/Conference Name, vol. X, no. Y, pp. 121–130, Year.

[14] Z. Author, “Soil type classification and mapping using hyperspectral remote sensing data,” Journal/Conference Name, vol. X, no. Y, pp. 131– 140, Year.

[15] A. B. Hossen, P. K. Diwakar, and S. Ragi, “Total nitrogen estimation in agricultural soils via aerial multispectral imaging and LIBS,” Journal/Conference Name, vol. X, no. Y, pp. 141–150, Year.

[16] C. Author and D. Author, “Use of clustering with partial least squares regression for predictions based on hyperspectral data,” Journal/Conference Name, vol. X, no. Y, pp. 151–160, Year.

[17] E. Author, “Use of laboratory hyperspectral reflectance data of soils for predicting their diurnal albedo dynamics accommodating their roughness,” Journal/Conference Name, vol. X, no. Y, pp. 161–170, Year.

[18] F. Author and G. Author, “An Integrated Framework for Predicting Soil Fertility and Crop Suitability Using Deep Learning,” Journal/Conference Name, vol. X, no. Y, pp. 171–180, Year.

[19] H. Author, I. Author, and J. Author, “Optimizing Fertilizer Usage with AI- Based Soil Nutrient Analysis,” Journal/Conference Name, vol. X, no. Y,

[20] K. Author and L. Author, “A Novel Machine Learning Model for Soil Moisture Prediction Using Remote Sensing Data,” Journal/Conference Name, vol. X, no. Y, pp. 191–200, Year.[21] M. Author, N. Author, and O. Author, “Automated Crop and Fertilizer Recommendation Using Sensor Data and AI,” Journal/Conference Name, vol. X, no. Y, pp. 201–210, Year.

[22] P. Author and Q. Author, “Soil Quality Prediction and Smart Farming: A Machine Learning Approach,” Journal/Conference Name, vol. X, no. Y, pp. 211–220, Year.

[23] C. Zhang, F. Liu, and Y. He, ‘‘Hyperspectral imaging-based soil organic matter prediction using deep learning,’’ Geoderma, vol. 387, p. 114929, 2021.

[24] X. Xie, J. Wang, and X. Chen, ‘‘Deep learning for hyperspectral image classification: An overview,’’ IEEE Transactions on Geoscience and Remote Sensing, vol. 58, no. 7, pp. 3866–3885, Jul. 2020.

[25] X. Zhao, C. Shi, and Y. Liu, ‘‘Combining CNN and attention mechanism for hyperspectral image classification,’’ Remote Sensing, vol. 13, no. 14, p. 2745, Jul. 2021.

[26] Z. Liu, T. Lin, and H. Luo, ‘‘Spectral-spatial classification of hyperspectral images using residual CNN and superpixel segmentation,’’ ISPRS Journal of Photogrammetry and Remote Sensing, vol. 149, pp. 70–84, Jan. 2019.

[27] D. Hong, N. Yokoya, and J. Chanussot, ‘‘More diverse means better: Multimodal deep learning meets remote sensing imagery classification,’’ IEEE Transactions on Geoscience and Remote Sensing, vol. 59, no. 3, pp. 4340–4354, Mar. 2020.

[28] Y. Zheng, X. Huang, and J. Li, ‘‘Multiscale convolutional neural network for soil property prediction using hyperspectral data,’’ Computers and Electronics in Agriculture, vol. 182, p. 106019, Sep. 2021.

[29] B. Liu, J. Sun, and Q. Zhou, ‘‘Hyperspectral imaging and machine learning for soil heavy metal detection: A review,’’ Environmental Pollution, vol. 300, p. 118986, Jun. 2022.

[30] C. He, Q. Wang, and F. Li, ‘‘A deep learning approach for estimating soil moisture using hyperspectral data,’’ Journal of Hydrology, vol. 577, p. 123949, Nov. 2019.

[31] W. Sun, C. Nie, and H. Zhang, ‘‘Soil texture classification using hyperspectral imaging and deep learning,’’ Remote Sensing of Environment, vol. 240, p. 111691, Mar. 2020.

[32] L. Zhang, L. Zhang, and B. Du, ‘‘Deep learning-based feature extraction for hyperspectral images: A review,’’ IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 14, pp. 3714–3731, 2021.

[33] H. Jiang, X. Wang, and Y. Feng, ‘‘Soil nutrient content estimation using deep learning and hyperspectral imaging,’’ Agricultural Systems, vol. 191, p. 103182, Feb. 2021.

[34] M. Li, J. Tang, and S. Yang, ‘‘Advances in hyperspectral image classification using deep learning models,’’ Neurocomputing, vol. 468, pp. 199–215, Oct. 2021.

[35] T. Chen, M. Lin, and R. Xu, ‘‘Fusion of hyperspectral and LiDAR data for improved soil property mapping using deep learning,’’ International Journal of Applied Earth Observation and Geoinformation, vol. 103, p. 102569, Dec. 2022.