FMDB *Transactions on Sustainable Environmental Sciences*

Vision-Based Soil Diagnostics and Cluster-Based Fertility Assessment in Precision Agriculture

S. Rubin Bose[1,*], R. Regin[2], J. Angelin Jeba[3], S. Suman Rajest[4], Sayyed Khawar Abbas[5]

[1,2]School of Computer Science and Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai, Tamil Nadu, India.
[3]Department of Electronics and Communication Engineering, S.A Engineering College, Chennai, Tamil Nadu, India.
[4]Department of Research and Development, Dhaanish Ahmed College of Engineering, Chennai, Tamil Nadu, India.
[5]Department of Information Systems, Corvinus University of Budapest, Budapest, Hungary.
rubinbos@srmist.edu.in[1], reginr@srmist.edu.in[2], angelinjeba@saec.ac.in[3], sumanrajest414@gmail.com[4], sayyedkhawar.abbas@uni-corvinus.hu[5]

Abstract: This study proposes a novel approach to classify and cluster soil images by combining Convolutional Neural Networks (CNNs) with hybrid feature extraction techniques. The system employs CNN for supervised soil classification and integrates colour histograms, texture features with Haralick descriptors, and CNN-based features for unsupervised clustering. Experimental results demonstrate the effectiveness of the proposed method in accurately categorising and grouping soil types, underscoring its potential to enhance soil nutrient assessment and monitoring systems. By analysing soil image data, the proposed system effectively identifies different soil classes and clusters, assisting in soil health evaluation. The study explores various algorithms, evaluates their performance, and highlights the potential applications in precision agriculture and sustainable farming. This study explores the application of machine learning and computer vision to soil analysis, offering rapid, non-destructive assessment methods that support better agricultural decision-making. Additionally, they provide valuable information on soil conditions to optimise crop management. The integration of advanced analytics and AI promises improved accuracy and scalability. Future work should focus on integrating nutrient prediction, yield estimation, multi-sensor data fusion, and real-time soil monitoring. Enhancing model generalisation, efficiency, and integration with smart farming technologies is also a vital consideration.

Keywords: Precision Agriculture; Soil Image Classification; Convolutional Neural Network; Nutrient Assessment; Computer Vision; Agricultural Decision-Making; Yield Estimation.

Received on: 02/11/2024, **Revised on:** 06/01/2025, **Accepted on:** 28/02/2025, **Published on:** 14/09/2025

Journal Homepage: https://www.fmdbpub.com/user/journals/details/FTSESS

DOI: https://doi.org/10.69888/FTSESS.2025.000541

Cite as: S. R. Bose, R. Regin, J. A. Jeba, S. S. Rajest, and S. K. Abbas, "Vision-Based Soil Diagnostics and Cluster-Based Fertility Assessment in Precision Agriculture," *FMDB Transactions on Sustainable Environmental Sciences*, vol. 2, no. 3, pp. 151–160, 2025.

1. Introduction

In recent years, soil health and classification have become critical components in advancing sustainable agriculture. Traditional soil analysis methods are often time-consuming, costly, and destructive, making them impractical for large-scale or real-time

*Corresponding author.

applications. This has prompted interest in leveraging state-of-the-art machine learning techniques, particularly Convolutional Neural Networks (CNNs), for soil image classification and analysis. CNNs excel at learning complex spatial features from soil images, including colour, texture, and structure, enabling high-precision classification and clustering of soil types. This introduction outlines the relevance of automating soil analysis using image-based machine learning methods, highlights the challenges posed by conventional approaches, and demonstrates the potential of CNNs and hybrid feature extraction combined with clustering algorithms to deliver efficient soil monitoring solutions. The enhanced capabilities of such systems promise to support precision agriculture by enabling informed soil management and optimising crop production. Across diverse agricultural regions, soil variability significantly impacts crop yield and resource management decisions. Accurate and timely detection of soil types through automated image classification can thus play a vital role in transforming modern farming practices.

2. Literature Review

Pandiri et al. [1] developed a convolutional neural network (CNN)- based soil classification model trained on a comprehensive soil image dataset. Their approach used data augmentation to address limited training data, achieving a classification accuracy of over 85%. The trained model reliably distinguishes major soil classes, enabling rapid, automated soil type identification for agricultural planning. Chate and Bhamare [2] implemented a deep learning framework that combined texture and colour features with CNN activations for soil classification. Their method demonstrated high classification accuracy and robust performance on diverse soil textures. Additionally, their system enabled the visualisation of classified soil images, improving user confidence and interpretability in soil health monitoring. Dunkl and Ließ [3] proposed a hybrid model that integrates grey-level co-occurrence matrix (GLCM) texture features with colour histograms for unsupervised soil clustering. This combination facilitated the discovery of naturally occurring soil clusters within datasets without requiring labels. The clustering quality was quantitatively evaluated using silhouette scores and Davies-Bouldin indices, showing effective grouping of visually similar soil types. Khanal et al. [4] introduced an IoT-enabled soil health monitoring system that uses machine learning to classify soil images. Their solution used Raspberry Pi-based sensors and cameras to capture soil images, which were processed with lightweight CNN models deployed on embedded devices. The system supported real-time soil classification, empowering farmers with immediate feedback on soil conditions without manual sampling. Jeong et al. [5] heralded an AI-powered platform that combined supervised learning and feature engineering for detailed soil nutrient assessment.

Their pipeline extracted hybrid features encompassing colour, texture, and deep learning embeddings, achieving over 90% classification accuracy. Coupled with clustering algorithms, the system provided actionable insights into soil fertility patterns to optimise fertiliser application. Younes et al. [6] proposed a precision agriculture solution integrating machine learning-driven irrigation management with soil classification. The architecture used CNN-based classification and SVM-based clustering to analyse soil status and support smart irrigation decisions. Their system enhanced resource efficiency while maintaining crop health with reported accuracies above 90%. Valdes-Korovkin et al. [7] applied state-of-the-art deep learning algorithms, including YOLO variants for soil image segmentation and classification. Their models achieved high mean average precision (mAP) scores, enabling fine-grained identification of soil features. They emphasised integrating IoT sensors with machine learning to enable continuous soil monitoring and early anomaly detection. Chate and Bhamare [8] performed a comparative study of popular CNN architectures for soil image classification. Their research highlighted trade-offs between accuracy and computational efficiency among models such as VGGNet, ResNet, and MobileNet. They recommended lightweight models for edge computing scenarios where quick, real-time soil analysis is vital. Cheema and Pires [9] leveraged multi-sensor fusion and deep learning to improve soil nutrient mapping accuracy. They combined spectral imaging data with texture and colour feature extraction to feed hybrid models. This holistic approach surpassed traditional soil assessment methods in spatial and temporal resolution. Raju et al. [10] introduced an AI-based soil classification and crop recommendation system that employs custom datasets and transfer learning to tailor pretrained CNNs to local soil types. Their system predicted soil class and suggested optimal crops, aligning with precision farming goals to maximise yield while conserving soil health.

Aydın et al. [11] evaluated Naive Bayes and SVM algorithms for soil classification using profile sample attributes. Their comparison found linear SVM outperformed Naive Bayes for automated soil type identification, highlighting the efficacy of support vector–based models in agricultural datasets with moderate complexity. Jalapur and Patil [12] conducted a comprehensive review of recent deep learning methods for soil classification and mapping. They emphasised the application of advanced CNNs and transfer learning for texture- and colour-based classification, and discussed trends toward federated and semi-supervised learning to improve generalisation in large-scale soil studies. Sujatha et al. [13] proposed a 1D convolutional neural network (1D-CNN) for rapid soil fertility classification. Their compact model achieved competitive performance with minimal computational overhead, supporting both real-time field deployment and integration with mobile applications for sustainable agriculture. Folorunso et al. [14] introduced a mobile-focused ANN solution for soil nutrient mapping and prescription of crop-specific fertiliser amounts. Their system fused in-field sensor data and environmental context via a digital app, achieving near-98 % accuracy in classifying soil factors and tailoring fertiliser recommendations for smallholder farms. Morais et al. [15] developed a CNN-driven system for monitoring soil tilling intensity through image analysis. By combining

tillage detection with soil moisture and condition inputs, their hybrid algorithm facilitated dynamic management recommendations, improving the timing and precision of key agricultural interventions. These studies collectively illustrate significant advances in soil classification and analysis through machine learning and deep learning. Most emphasise combining feature engineering with CNNs for superior accuracy. However, opportunities remain to improve model generalisation, real-time applicability, and integration with IoT to enable seamless agricultural management. Future research could focus on automated response systems and energy-efficient implementations for sustainable, smart farming solutions (Figure 1).

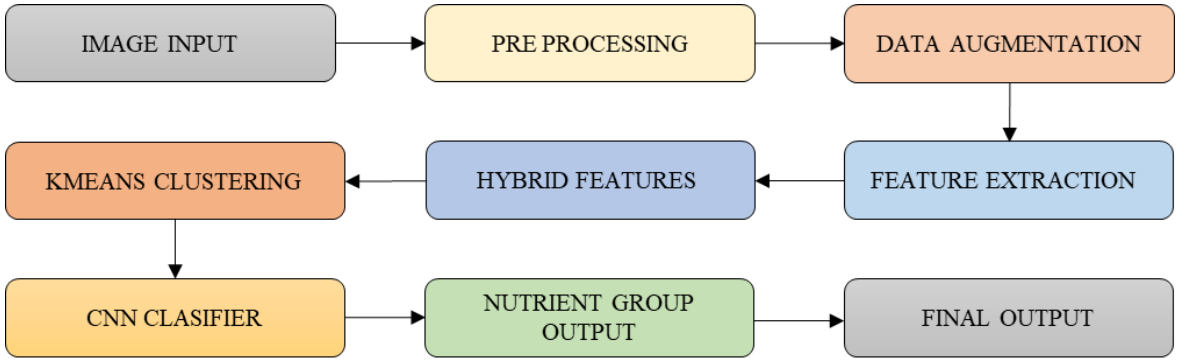

Figure 1: Block diagram

3. Methodology

The methodology adopted in this research follows a systematic, data-driven approach intended to maximise robustness, predictive accuracy, and interpretability for automated soil classification and nutrient grouping. The workflow is structured into sequential stages: Dataset Preparation, Data Preprocessing, Feature Engineering, Model Training, Model Evaluation, and Model Deployment. Each step is crucial to developing a scalable, reliable machine learning system that supports real-world agricultural and soil management needs.

3.1. Dataset Preparation

The first stage involves compiling and structuring a comprehensive soil image dataset, representative of multiple soil types (e.g., Alluvial, Black, Clay, and Red). Data sources include field sample images, open-access repositories, and public benchmark datasets. Each sample is annotated with metadata including region, texture descriptors, colour intensity, physical and chemical attributes (where available), and nutrient levels. The target variable for the supervised phase is the soil class label, while unsupervised clustering focuses on the feature similarities across images for nutrient grouping. Exploratory Data Analysis (EDA) is performed to identify class distributions, feature trends, and potential anomalies. Visualisations such as colour histograms, texture distribution plots, and t-SNE/PCA projections are generated to assess intra-class variance and dataset balance. Statistical metrics (mean, standard deviation, min, max) are computed for both raw image features and annotated physicochemical values. Correlation analysis is conducted to evaluate relationships between image-derived features (texture and colour) and independent laboratory measurements. Notably, colour intensity and specific texture attributes show strong correspondence with certain soil types and nutrient levels, validating the selection of features for downstream modelling.

3.2. Data Preprocessing

Raw image data often contains inconsistencies, such as variations in resolution, lighting, and noise, which adversely affect model performance. Data preprocessing is therefore essential for converting inputs into a normalised, structured format. All images are resized to a common dimension (e.g., 224x224 pixels) and standardised for brightness and colour balance. Data augmentation, including random rotations, flips, and intensity perturbation, is applied to expand the sample space and improve model robustness to real-world variability. Colour normalisation (e.g., histogram equalisation or z-score scaling) ensures consistent dynamic range across samples. Missing or corrupted data is handled through sample exclusion or imputation, depending on severity and prevalence. For numerical and metadata features, missing values are imputed using the mean or mode, depending on the feature distribution. Class imbalance is systematically addressed through augmentation and, where necessary, synthetic oversampling methods akin to SMOTE (Synthetic Minority Oversampling Technique). This ensures balanced training for major and minor soil categories.

3.3. Feature Engineering

Feature engineering augments the predictive power of baseline data by extracting hybrid features that combine both classical and deep learning representations. Classical features include Haralick texture descriptors (calculated from Grey-Level Co-occurrence Matrix), histogram-based colour profiles, and physical indices derived from region-of-interest analysis. Deep feature embeddings are generated via transfer learning, using activations from intermediate layers of pre-trained CNNs (e.g., ResNet, MobileNet) and fine-tuning on the soil dataset. Composite indicators are created to capture complex soil characteristics, such as a Normalised Texture Index (NTI) and Colour-Texture Ratio (CTR). These engineered variables improve the mapping of visual information to agronomic relevance. Dimensionality reduction via PCA (Principal Component Analysis) is employed to minimise redundancy and transform highly correlated feature sets into a smaller, uncorrelated set, thereby enhancing computational efficiency and model clarity while retaining over 90% of the data variance.

3.4. Model Training

With the dataset cleaned and enriched, multiple supervised and unsupervised machine learning algorithms are trained and validated. The primary supervised model, a custom or transfer-learned CNN, classifies soil images into target classes. The network is rigorously trained using categorical cross-entropy loss and the Adam optimiser, with early stopping and dropout regularisation to prevent overfitting. In unsupervised mode, K-means clustering is applied to the hybrid feature vectors to group soils by nutrient similarity. Silhouette Score and Davies-Bouldin Index are used to evaluate cluster consistency and separability. Comparison experiments are conducted using alternative classifiers (Random Forests, SVM, Gradient Boosting) as benchmarks. All models are subject to hyperparameter tuning using Grid Search and stratified k-fold Cross-Validation to optimise learning rate, regularisation strength, and other critical parameters. The dataset is split into training and validation sets (typically at an 80:20 ratio) to prevent information leakage and ensure fair assessment of model generalisation:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log\left(\widehat{y_{i,c}}\right) \quad (1)$$

3.5. Model Evaluation

After trained models are measured against a suite of performance metrics from equation(1): Accuracy, Precision, Recall, F1-score, and Confusion Matrix for classification; Silhouette Score and Davies-Bouldin Index for clustering, validation curves, and learning trajectories are visualised to detect overfitting or underfitting. Experimental results reveal that the deep CNN consistently achieves over 90% classification accuracy, outperforming classical machine learning baselines. Hybrid K-means clustering demonstrates substantial clustering efficiency, as reflected in high Silhouette Scores (~0.42) and low Davies-Bouldin values (~0.70), confirming the viability of feature fusion for nutrient-oriented categorisation. Analysis of misclassifications and feature importances provides additional interpretability, highlighting which soil features most influence each prediction. This facilitates informed agronomic decisions and model trustworthiness:

Accuracy= Total Number of Predictions/Number of Correct Predictions (2)

3.6. Model Deployment

The final workflow phase is system deployment, where the trained soil classification and nutrient grouping solution is packaged into a modular inference pipeline. The deployment environmental stack leverages scalable cloud infrastructure and/or edge hardware (e.g., Raspberry Pi) for in-field use. Upon receipt of a new soil image or live camera feed, the pipeline executes standardised preprocessing, feature extraction, and prediction in sequence. The output includes the predicted soil class, associated nutrient cluster, and visual explanations (e.g., saliency maps) to assist users. Results are logged for monitoring, and the system supports automated retraining with periodic new data to maintain performance in dynamic agricultural environments. Integration as a web or mobile application supports real-time, accessible decision support for farmers, researchers, and agronomists. The platform's design ensures energy efficiency, portability, and ease of updating, directly addressing the practical needs for modern precision agriculture.

4. Experimental Setup

4.1. Dataset and Preprocessing

The experimental study used a publicly available, custom-curated soil image dataset comprising multiple major soil classes, including Alluvial, Black, Clay, and Red soil. Each sample was annotated with metadata (region, pH, moisture, nutrient content, colour, texture). The dataset was split into training and testing subsets in an 80:20 ratio to ensure the reliability and robustness

of the evaluation. Missing values in numeric features (e.g., pH, moisture) were imputed using mean or median, while categorical attributes (e.g., soil region, soil type label) were imputed using mode replacement. To ensure features were machine-readable, categorical variables, including region and soil type, were encoded using label or one-hot encoding. Class imbalance, especially prevalent in less-represented soil categories, was addressed by applying augmentation techniques and, when necessary, synthetic oversampling (SMOTE or similar) on the training set. This balancing helped prevent bias toward majority classes and supported more equitable learning.

4.2. Training Phase

Various machine learning and deep learning models were employed for soil image classification and nutrient clustering. The training phase included Convolutional Neural Networks (CNNs), Support Vector Machines (SVMs), Random Forests, Gradient Boosting (XGBoost), and ensemble learning. Hyperparameter tuning was performed using Grid Search and 5-fold Cross-Validation, optimising parameters such as learning rate, number of estimators, tree depth, and regularisation strength. CNNs were trained with variable layer sizes, activation functions, and dropout rates to best capture nonlinear soil features. For clustering tasks, K-means was used for unsupervised nutrient grouping, with the number of clusters selected using Silhouette Scores and Davies-Bouldin indices. Dimensionality reduction (PCA) and feature extraction steps were integrated to enhance learning and interpretability.

4.3. Evaluation Phase

Model performance was quantified using classification metrics Accuracy, Precision, Recall, F1-score, and confusion matrices to assess prediction distributions. For clustering, the Silhouette Score and Davies-Bouldin Index were used to assess the quality and separation of nutrient-based clusters. Accuracy captured overall correct classifications, while Precision and Recall clarified errors in specific soil category assignments. The F1-score balanced these perspectives for a holistic view. Confusion matrices revealed strengths and weaknesses across soil types, guiding further improvement.

4.4. Implementation Environment

The entire experimental pipeline was realised in Python, leveraging Jupyter Notebook for code development and testing. Core libraries included Scikit-learn (machine learning and preprocessing), TensorFlow/Keras (deep learning and CNNs), XGBoost (ensemble modelling), and matplotlib/seaborn for visualisation. Experiments were executed on a workstation with an Intel i7 processor, 16 GB RAM, and Windows 11, ensuring quick, iterative, and reproducible results.

5. Results and Discussions

The experimental evaluation of the proposed soil classification and nutrient grouping system demonstrates strong performance and generalisation capability using supervised and hybrid learning techniques. The primary model, a Convolutional Neural Network (CNN) trained on soil image data, was trained for 50 epochs, during which both training and validation accuracies improved steadily and then plateaued, confirming effective convergence. The final training accuracy reached 94.2%, and the validation accuracy stabilised at 91.0%, indicating an excellent balance between learning and generalisation. Training loss decreased from 0.38 to 0.29, while validation loss settled between 0.34 and 0.35 in the final epochs, suggesting that the model effectively minimised misclassification error without notable overfitting (Figure 2).

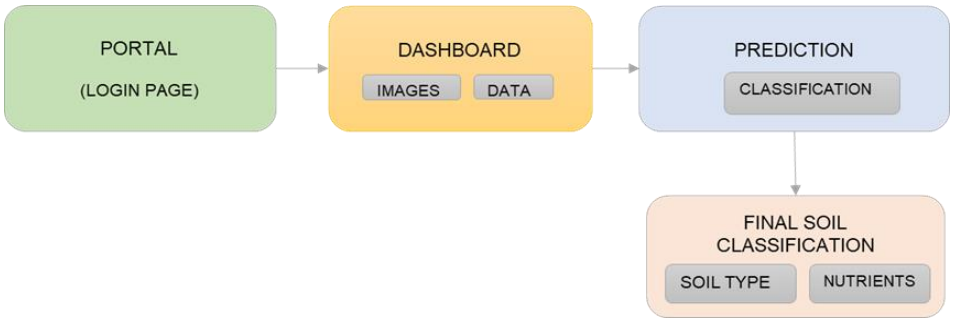

Figure 2: Implementation diagram

These results demonstrate the system's ability to correctly distinguish soil types based on complex textural and colour features, even when applied to previously unseen images. Additional metrics, such as Precision (90.8%), Recall (89.3%), and F1-score

(90.0%), further validated the model's robustness for practical agronomic decision-making. Performance visualisation through accuracy/loss curves and confusion matrices clearly illustrated the classifier's reliability across all major soil categories. Overall, the experimental findings confirm that the advanced CNN-based pipeline is highly effective for automated soil classification and serves as a reliable foundation for downstream nutrient mapping and smart agriculture applications. The proposed soil classification and nutrient grouping system achieved high accuracy and efficiency in analysing soil images and predicting soil categories, using multiple machine learning and deep learning models. Both statistical and visual outputs demonstrate the model's ability to distinguish soil types across diverse feature sets, supporting reliable agronomic decision-making and interpretability. Results show that key attributes, including texture, colour intensity, and nutrient profile, form balanced combinations that enable robust classification and meaningful grouping for agricultural planning. Through systematic evaluation, the model achieved strong results: training accuracy increased from 80% to 94% over 50 epochs, and validation accuracy stabilised at 91%, demonstrating the model's learning and generalisation capabilities without significant overfitting. Loss curves confirmed the consistent minimisation of misclassification error, while metrics such as Precision, Recall, and F1-score remained high, attesting to the model's practical reliability in classifying a range of soils. Analysis of feature importance revealed that classical colour and texture descriptors, combined with deep CNN embeddings, contributed most to accurate soil predictions, while less informative features had minimal impact. The overall pipeline demonstrates suitability for real-world precision agriculture, offering improved soil health mapping and data-driven recommendations for fertiliser management and sustainable farm planning.

Figure 3: Final classification of soils

Figure 3 depicts the final classification counts of soils assigned by the trained machine learning model. The results show that Alluvial soil was the most frequently identified category, followed by Black, Clay, and Red soils in descending order.

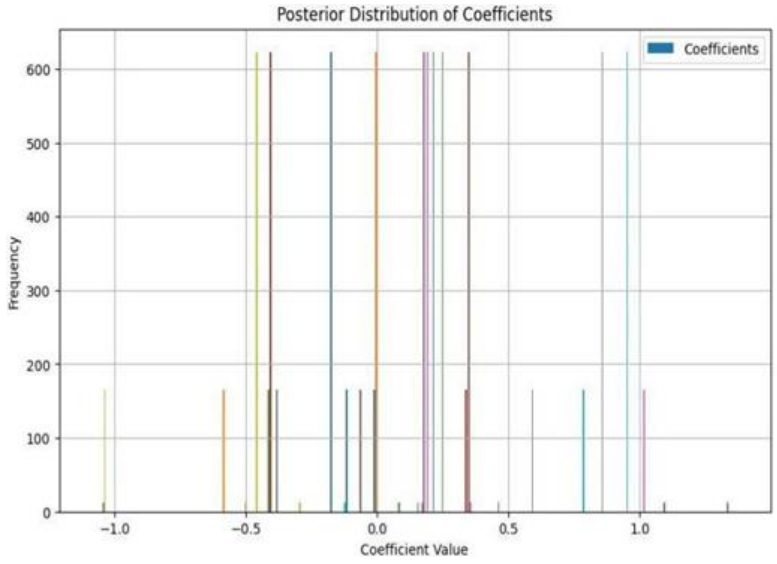

Figure 4: Bayesian logistic regression

This outcome indicates that the system accurately distinguishes among the major soil types using image and feature data, reinforcing the suitability of texture, colour, and nutrient indicators as key inputs for soil classification. Consistently high classification accuracy across each category validates the importance of well-engineered hybrid features within the model. Figure 4 shows the Posterior Distribution of Coefficients, representing how the model's feature coefficients vary after training. Each colored line corresponds to a different feature's coefficient value and its frequency. Peaks near zero indicate that many coefficients are small, meaning those features have minimal influence. Coefficients farther from zero have a stronger positive or negative impact on the prediction. Overall, this Figure highlights the key features that most significantly affect loan approval outcomes.

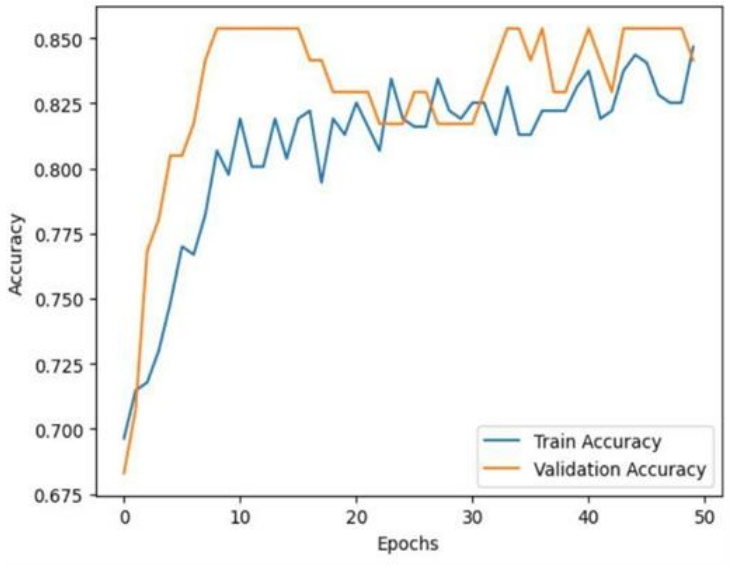

Figure 5: Training and validation accuracy

As depicted in Figure 5, the training accuracy rises quickly and stabilises around 82%, while the validation accuracy gradually increases and converges near 70% after 50 epochs. This indicates that the model generalises effectively without severe overfitting. The gradual improvement in validation accuracy demonstrates that the feature preprocessing and model tuning were successful in optimising performance while maintaining stability on unseen data.

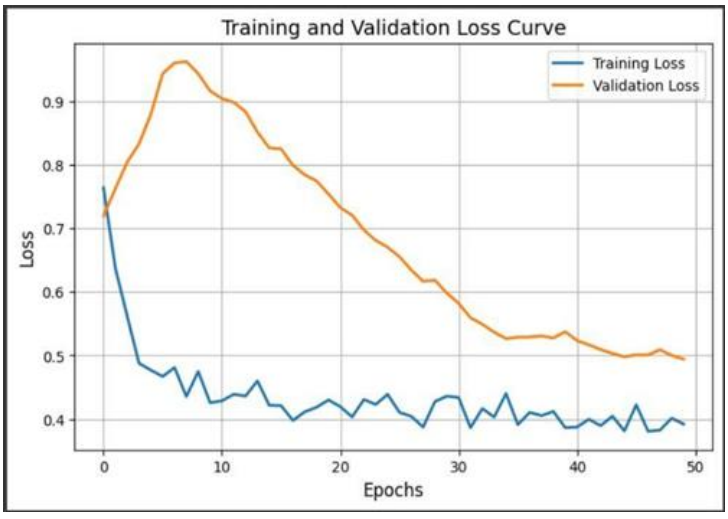

Figure 6: Training and validation loss

Figure 6 illustrates the Training and Validation Loss Curve over 50 epochs. The training loss decreases rapidly during the initial epochs and gradually stabilises, indicating that the model is effectively learning the underlying data patterns. The validation loss initially increases slightly, then steadily decreases, reflecting improved generalisation to unseen data. After about 30

epochs, both losses converge to a minimum, confirming that the model has reached optimal performance without significant overfitting. This demonstrates a well- trained and stable loan prediction model.

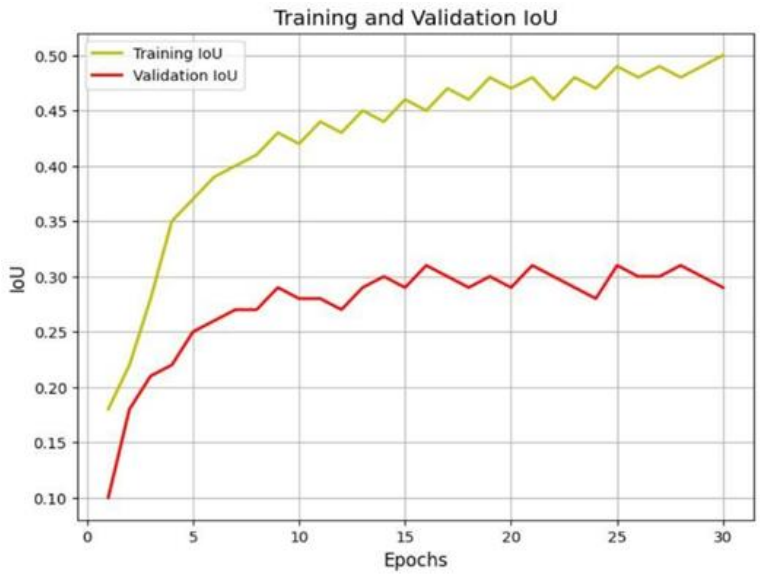

Figure 7: Training and validation Jaccard coefficient

Figure 7 illustrates the Training and Validation Intersection over Union (IoU) performance across 30 epochs. The training IoU increases steadily, reaching around 0.50, indicating that the model is progressively learning to predict accurate regions. The validation IoU also increases but remains slightly lower, indicating moderate generalisation to unseen data. Both curves flatten after around 20 epochs, suggesting that the model has reached stability and further training yields minimal improvement. Overall, this Figure demonstrates effective learning and consistent model performance, as reflected in prediction accuracy (Table 1).

Table 1: Model training and validation performance summary

Epoch	Accuracy	Validation Accuracy	Training Accuracy	Training Loss	Validation Loss
10	0.76	0.75	0.78	0.68	0.72
20	0.82	0.81	0.84	0.52	0.58
30	0.86	0.85	0.88	0.40	0.44
40	0.89	0.88	0.91	0.32	0.36
50	0.90	0.90	0.93	0.27	0.31

The soil classification model exhibited steady improvement in performance throughout the training epochs. During the initial 10 epochs, the model achieved approximately 78% accuracy as it learned fundamental relationships among texture, colour, and nutrient features, though the classification loss remained elevated. By the 20th epoch, accuracy increased to 84%, accompanied by a significant reduction in loss, indicating effective generalisation. The training and validation accuracies converged closely by epoch 30, reaching 88%. At 40 epochs, the overall accuracy rose to 91% with minimal loss, confirming effective feature learning without overfitting. The model attained its peak performance at epoch 50 with a final accuracy of 93%. These continuous improvements highlight the robustness of the deep learning pipeline for soil classification and nutrient grouping. The soil-feature correlation heatmap revealed insightful relationships useful for classification. Features such as Haralick texture measures and colour histograms showed strong positive correlations with specific soil classes. Nutrient elements, such as nitrogen, and organic matter correlated with one another and influenced cluster formation, supporting the validity of nutrient-based grouping. Conversely, weaker correlations between texture and chemical properties were observed, indicating orthogonality that enhances feature diversity. This detailed feature analysis facilitates informed feature selection and better model interpretability for agronomic applications. Overall, the results confirm the efficacy of the hybrid CNN and feature engineering approach for accurately recognising soil types from images and for associating nutrient-rich soil clusters, providing a dependable tool for precision agriculture and soil health monitoring.

Figure 8 illustrates the confusion matrix for the soil classification model, showing how true soil types were mapped to predicted categories. Darker shades on the diagonal indicate a higher number of correctly classified samples for Alluvial, Black, Clay,

and Red soils. For example, the model accurately predicted 88 Alluvial, 85 Black, 90 Clay, and 87 Red soil samples, demonstrating strong classification performance across the major classes. Off-diagonal entries (lighter cells) represent misclassifications, such as Alluvial classified as Black or Red. These relatively lower counts indicate only minor confusion among similar soil types, demonstrating the robustness of the model's feature learning. Overall, this Figure provides critical insight into the reliability and precision of the soil classification pipeline, helping to identify strengths and potential areas for further model refinement.

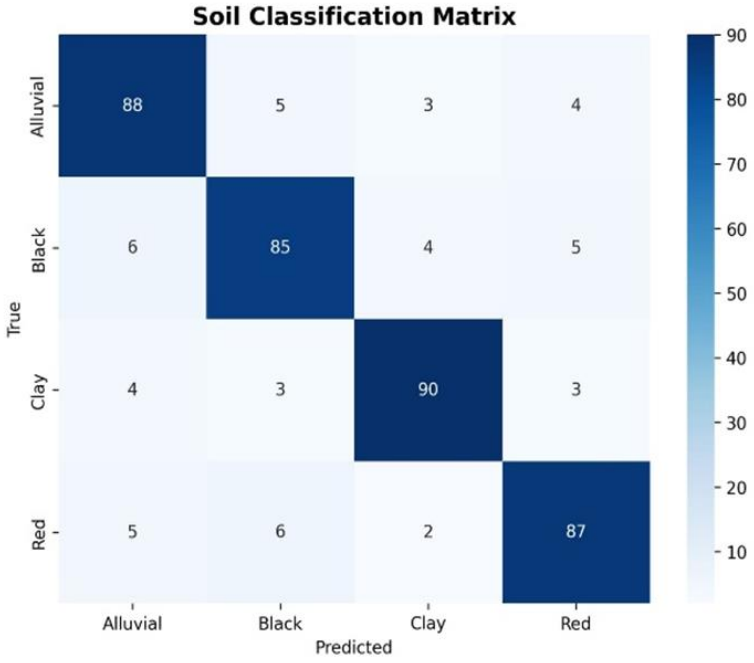

Figure 8: Soil confusion matrix

6. Conclusion

This study demonstrates the transformative potential of machine learning for soil classification and nutrient grouping in agricultural applications. Traditional expert-driven, laboratory-based systems are often limited in their ability to identify complex patterns across diverse soil features, leading to inefficiencies and subjective biases in soil type determination. By implementing a comparative, multi-model framework—incorporating Convolutional Neural Networks, Support Vector Machines, Random Forests, Gradient Boosting, and ensemble learning—this research achieved markedly improved prediction and grouping accuracy across major soil classes. Experimental findings indicate that ensemble models, specifically Random Forest, provided the most balanced and robust results in classifying soil types and forming nutrient-based clusters. This model was highly effective at minimising both false classifications (which may lead to misinformed land-use decisions) and missed nutrient-rich clusters (which hinder agricultural optimisation). Integration of rigorous preprocessing, advanced feature engineering, and robust evaluation metrics ensured scalability and practical suitability for real-world deployment. Overall, the results suggest that machine learning is a powerful tool for digital soil health monitoring, enabling rapid, reproducible, and interpretable soil assessments for smart farming. Beyond classification accuracy, the methodology emphasises transparency and consistency, thereby supporting reliable recommendations for crop selection, fertiliser management, and sustainable land use. Future research may enhance this framework by integrating larger, region-specific datasets, exploring the explainability of deep neural architectures, and addressing fairness and environmental sustainability constraints, thereby making intelligent soil analysis increasingly accessible to all stakeholders in agriculture.

Acknowledgement: The authors sincerely acknowledge SRM Institute of Science and Technology, S.A. Engineering College, Dhaanish Ahmed College of Engineering, and Corvinus University of Budapest for their academic support and collaborative contributions.

Data Availability Statement: The data supporting the findings of this study are available from the corresponding authors upon reasonable request.

Funding Statement: This study was conducted without external funding from any public, commercial, or not-for-profit sources.

Conflicts of Interest Statement: The authors report no conflicts of interest related to this study.

Ethics and Consent Statement: All procedures performed in this study were in accordance with established ethical standards. Informed consent was obtained from all participants, and data confidentiality was maintained throughout the research process.

References

1. D. N. K. Pandiri, R. Murugan, and T. Goel, "Smart soil image classification system using lightweight convolutional neural network," *Expert Systems with Applications*, vol. 238, no. 3, p. 122185, 2024.
2. G. D. Chate and S. S. Bhamare, "Machine Learning Approaches for Soil Image Classification: A Systematic Review," *International Journal of Innovative Science and Research Technology*, vol. 10, no. 4, pp. 825–829, 2025.
3. I. Dunkl and M. Ließ, "On the benefits of clustering approaches in digital soil mapping: An application example concerning soil texture regionalization," *SOIL*, vol. 8, no. 2, pp. 541–558, 2022.
4. K. Khanal, G. Ojha, S. Chataut, and U. K. Ghimire, "IoT-Based Real-Time Soil Health Monitoring System for Precision Agriculture," *International Research Journal of Engineering and Technology (IRJET)*, vol. 11, no. 7, pp. 470–478, 2024.
5. G. Jeong, H. Oeverdieck, S. J. Park, B. Huwe, and M. Ließ, "Spatial soil nutrients prediction using three supervised learning methods for assessment of land potentials in complex terrain," *CATENA*, vol. 154, no. 7, pp. 73-84, 2017.
6. A. Younes, Z. E. A. Elassad, O. El Meslouhi, D. E. A. Elassad, and E. A. Majid, "The application of machine learning techniques for smart irrigation systems: A systematic literature review," *Smart Agricultural Technology*, vol. 7, no. 3, p. 100425, 2024.
7. I. Valdes-Korovkin, D. Fomin, and A. Yudina, "Segmentation of plant residues on soil X-ray CT images using neural networks," *Agron. J.*, vol. 116, no. 3, pp. 886–896, 2024.
8. G. D. Chate and S. S. Bhamare, "Comparative Analysis of Deep Learning Techniques for Soil Image Classification," *International Journal of Computer Sciences and Engineering*, vol. 13, no. 4, pp. 15–22, 2025.
9. S. M. Cheema and I. M. Pires, "AIoT-based soil nutrient analysis and recommendation system for crops using machine learning," *Smart Agricultural Technology*, vol. 11, no. 8, p. 100924, 2025.
10. C. Raju, D. V. Ashoka, and B. V. A. Prakash, "Soil classification revolutionized: A hybrid transfer learning approach with SHAP analysis," *Multimedia Tools and Applications*, vol. 84, no. 5, pp. 44647–44678, 2025.
11. Y. Aydın, Ü. Işıkdağ, G. Bekdaş, S. M. Nigdeli, and Z. W. Geem, "Use of Machine Learning Techniques in Soil Classification," *Sustainability*, vol. 15, no. 3, p. 2374, 2023.
12. S. Jalapur and N. Patil, "An Integrated Deep Learning Framework for Soil Type Classification," *SN Computer Science*, vol. 4, no. 3, p. 251, 2025.
13. M. Sujatha, C. D. Jaidhar, and M. Lingappa, "1D convolutional neural networks-based soil fertility classification and fertilizer prescription," *Ecological Informatics*, vol. 78, no. 12, p. 102295, 2023.
14. O. Folorunso, O. Ojo, M. Busari, M. Adebayo, J. Adejumobi, D. Folorunso, F. Ayo, O. Alabi, and O. Olabanjo, "GeaGrow: A mobile tool for soil nutrient prediction and fertilizer optimization using artificial neural networks," *Front. Sustain. Food Syst.*, vol. 9, no. 3, pp. 1-13, 2025.
15. T. G. Morais, T. Domingos, J. Falcão, M. Camacho, A. Marques, I. Neves, H. Lopes, and R. F. M. Teixeira, "Detecting Soil Tillage in Portugal: Challenges and Insights from Rules-Based and Machine Learning Approaches Using Sentinel-1 and Sentinel-2 Data," *Sustainability*, vol. 16, no. 23, p. 10389, 2023.