CoreID JOURNAL

Vol. 3 No. 2, July 2025 (60-64)
ISSN: 2987-6990
DOI: https://doi.org/10.60005/coreid.v3i2.142

Convolutional Neural Network for Soil Surface Image Classification in Six Soil Categories

Muhamad Iqbal Fauzi[1], Nasywah Darraini[2], Muhamad Randi Septiansah[3], Erwinestri Hanidar Nur Afif[4]
[1,2,3]Informatics Engineering, Faculty of Science and Technology, UIN Sunan Gunung Djati Bandung, Indonesia
[4]IAIN Sorong, Indonesia

Article Info

Article history:

Received April 11, 2025
Revised May 12, 2025
Accepted July 12, 2025

Keywords:

Convolutional Neural Network
soil image classification
soil type

ABSTRACT

Soil type classification is important for agriculture, geology, and civil engineering because soil characteristics influence land suitability, tillage strategy, irrigation, fertilization, and foundation stability. However, manual soil identification through field observation or laboratory analysis can be time-consuming and may introduce subjective errors. This study proposes an automated soil image classification approach using a Convolutional Neural Network (CNN). The dataset comprises six soil categories-black soil (tanah hitam), yellow soil (tanah kuning), peat soil (tanah gambut), cinder/volcanic soil (tanah vulkanik), laterite soil (tanah laterit), and cracked soil (tanah retak) -collected from a public Kaggle dataset and complemented with web-extracted cracked-soil images. Images are preprocessed through resizing, normalization, and training-time augmentation before being split into training, validation, and testing subsets. Experimental results show that the proposed CNN achieves 91.61% test accuracy and substantially improves performance compared to training without preprocessing. These findings indicate that CNN-based models, supported by appropriate preprocessing, can provide practical decision support for rapid soil type identification under diverse image conditions.

Corresponding Author:

Muhamad Iqbal Fauzi
Informatics Engineering, Faculty of Science and Technology, UIN Sunan Gunung Djati Bandung
Jl. A.H. Nasution No. 105 Cibiru, Kota Bandung, Jawa Barat 40614, Indonesia
Email: miqbalfauzi000@gmail.com

1. INTRODUCTION

Soil is a fundamental natural resource that directly determines agricultural productivity, land suitability, and sustainable crop management. In real field practice, farmers and agricultural practitioners frequently face heterogeneous soil conditions even within nearby plots, causing differences in water retention, nutrient availability, and tillage requirements [1], [2], [3], [4]. Accurate soil-type identification therefore becomes a practical need to support decisions such as crop selection, irrigation scheduling, fertilization strategy, and land preparation. However, conventional identification that relies on expert observation or laboratory analysis is often time-consuming, requires specific expertise and equipment, and may not be feasible for rapid decision-making in the field [5], [6]. With the increasing availability of cameras and smartphones, soil images offer a practical pathway for faster identification and decision support through automated classification systems.

Research on soil-type identification and classification can be grouped into at least three streams. First, studies in soil science and agronomy emphasize field morphology and laboratory-based characterization (e.g., texture, organic content, moisture, and color descriptors) that provide strong scientific grounding but often require controlled procedures and specialized tools, limiting rapid deployment in everyday practice [6]. Second, machine learning approaches such as K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Naive Bayes have been applied to classification tasks, commonly by extracting handcrafted features (e.g., color histograms or texture descriptors) and training conventional classifiers [6], [7], [8], [9], [10], [11], [12]. While these methods can perform reasonably under controlled settings, they may degrade under real-world variations such as lighting, camera angle, soil moisture, shadows, and mixed particles, because the features are manually engineered and not always robust. Third, deep learning—particularly Convolutional Neural Networks (CNN)—has become a dominant approach for image classification, including agriculture-related tasks, because it learns hierarchical features directly from images and often improves generalization on complex visual patterns [6], [9], [13]. Despite this progress, important gaps remain: many studies focus on limited class diversity, do not explicitly target locally relevant soil categories, or lack thorough evaluation of robustness under practical image variation. In addition, datasets for multi-class soil recognition are often small and heterogeneous, and many works do not benchmark CNN performance against simpler baselines in a comparable setup. These limitations motivate a more structured investigation of CNN-based soil classification for multiple soil types using image data captured in practical conditions.

To address the above gaps, this study aims to develop and evaluate a CNN-based soil image classification model for six soil categories commonly referenced in local field contexts: black soil (tanah hitam), yellow soil (tanah kuning), peat soil (tanah gambut), cinder/volcanic soil (tanah vulkanik), laterite soil (tanah laterit), and cracked soil (tanah retak). Specifically, this research (i) compiles a labeled image dataset for the six classes, (ii) applies a preprocessing pipeline and data augmentation to improve robustness to visual variation, and (iii) evaluates the CNN model using standard performance metrics (accuracy, precision, recall, F1-score, and confusion matrix). Where appropriate, CNN performance is compared with conventional baselines to highlight the value of deep feature learning for soil images .

This research is grounded in the argument that soil-type recognition is primarily a visual-pattern problem—driven by color distribution, surface texture, granularity, and structural cues (e.g., cracking patterns)—which can be learned effectively by CNN architectures through end-to-end training. Accordingly, the main hypothesis is that a CNN-based model will achieve higher and more stable classification performance than conventional approaches that rely on handcrafted features, particularly under varying lighting and capture conditions [6], [8], [9], [13]. A secondary hypothesis is that appropriate preprocessing and augmentation will improve generalization by reducing sensitivity to common noise factors (illumination changes, shadows, and minor viewpoint shifts). Finally, it is expected that misclassifications will concentrate among visually similar categories (e.g., black vs. peat under low light, or yellow vs. laterite when color saturation is similar), and that analyzing these confusion patterns can guide data collection improvements and model refinement.

To strengthen the theoretical foundation of this study, the following review highlights key concepts related to soil types and visual characteristics, image preprocessing and augmentation, and CNN-based image classification.

2. METHOD

2.1 Unit of Analysis

The unit of analysis in this study is individual soil-surface images. Each image constitutes one observation and is labeled into one of six soil categories: black soil, yellow soil, peat soil, cinder/volcanic soil, laterite soil, and cracked soil. The classification task aims to learn visual soil cues such as color distribution, granularity, surface texture, and cracking patterns from RGB images.

2.2 Research Design

This study employs a quantitative experimental design within a supervised multi-class image classification framework. The experimental approach is used to systematically evaluate model performance under different architectural and training settings. CNN is selected because soil type recognition is strongly driven by complex visual patterns; CNN can learn hierarchical features directly from images and typically outperforms approaches that rely on handcrafted features under varying acquisition conditions [1], [2], [3], [14].

2.3 Data Sources

Two data sources are used. First, a public Kaggle dataset provides 300 RGB JPG soil images distributed across five classes (black, yellow, peat, laterite, and cinder/volcanic) [15]. Second, approximately 50 cracked-soil images are collected through web-image extraction and manual screening to complete the six-class scope. Overall, the compiled dataset contains approximately 350 labeled images.

Table 1 presents representative samples from the soil image dataset used in this study. The dataset consists of RGB soil-surface photographs in JPG format that are grouped into five main categories from the

Kaggle source-black soil, cinder soil, laterite soil, peat soil, and yellow soil-which are visually distinguishable through differences in dominant color, granularity, and surface texture. Each class exhibits characteristic appearance patterns, for instance darker tones in black and peat soil, granular and coarse particles in cinder soil, reddish-brown hues in laterite soil, and lighter yellowish coloration in yellow soil. These sample images are provided to illustrate the visual variability within the dataset and to support the rationale for using CNN-based feature learning in the subsequent classification experiments.

Table 1. Dataset

No	Image	Soil type
1		Black soil
2		Cinder soil
3		Laterite soil
4		Peat soil
5		Yellow soil
6		Tanah retak

2.4 Data Collection and Preparation

Kaggle images are downloaded and organized by class based on the provided labels. Cracked-soil images are retrieved via keyword-based web search and curated by removing non-soil images, duplicates, and samples with excessive background clutter. As shown in Figure 1, preprocessing includes resizing images to a fixed input dimension, normalizing pixel values, and applying training-time augmentation (rotation, shifting, shearing, zooming, and flipping) to increase robustness and reduce overfitting. The dataset is split into training, validation, and testing subsets using stratification to preserve class proportions.

2.5 Data Analysis

Data analysis is conducted by training CNN models and evaluating their classification performance. The CNN architecture (Figure 2) consists of convolutional layers, max-pooling layers, fully connected layers, and dropout regularization. Experiments vary (i) convolution and pooling configurations, (ii) dropout rate, (iii) optimizer (SGD, RMSProp, Adam), and (iv) learning rate values to identify the best-performing model. Model selection is based on validation performance, while final reporting uses the held-out test set. Evaluation emphasizes accuracy together with class-sensitive metrics (precision, recall, F1-score) and a confusion matrix to analyze misclassifications.

3. RESULTS AND DISCUSSION

3.1 Dataset Characteristics

The compiled dataset contains approximately 350 soil images across six classes. The Kaggle portion contributes 300 images across five classes, and the cracked-soil class is complemented with around 50 web-extracted images. Table 1 summarizes the dataset composition.

Table 1. Dataset composition

Class	Source	Approx. count
Black soil	Kaggle	≈60
Cinder/volcanic soil	Kaggle	≈60
Laterite soil	Kaggle	≈60
Peat soil	Kaggle	≈60
Yellow soil	Kaggle	≈60
Cracked soil	Web extraction	≈50

3.2. Experimental Setup and Training Behavior

Experiments were conducted by modifying the CNN configuration (number of convolution and pooling layers), adjusting dropout rates, and comparing optimizers (SGD, RMSProp, Adam) under different learning rates. Figure 3 (to be inserted) is recommended to present training and validation accuracy/loss curves for the selected best model. Overall, models trained with preprocessing and augmentation converged more stably and achieved higher validation performance than models trained without these steps.

3.3. Classification Performance and Error Analysis

The best-performing CNN model achieved 91.61% accuracy on the test set. This result indicates that the model can learn discriminative soil features and generalize to unseen images. To provide deeper insight beyond accuracy, Figure 4 (to be inserted) should present a confusion matrix, enabling identification of classes that are frequently confused.

Qualitative error analysis suggests that misclassifications occur mainly between visually similar categories. For example, cracked-soil images with dark coloration may be predicted as black soil, while yellowish cracked-soil images may be predicted as yellow soil. These errors are consistent with the hypothesis that color dominance can override structural cues when cracking patterns are subtle or obscured by lighting conditions

3.4. Discussion

The results demonstrate that CNN-based classification can effectively distinguish multiple soil types from images, achieving high test accuracy in a multi-class setting. The substantial improvement observed after applying preprocessing and augmentation supports prior findings that augmentation can mitigate limited data and improve generalization in image classification. Practically, the proposed approach can reduce reliance on time-consuming manual identification and provide rapid decision support for soil-related agricultural practices.

Nevertheless, the study reveals challenges in separating classes with overlapping visual characteristics. Soil appearance is influenced by acquisition conditions such as illumination, moisture, camera distance, and background clutter. In particular, color similarity between peat and black soil, or between yellow and laterite soil, may increase confusion when texture cues are weak. Future work should therefore consider expanding the dataset, ensuring balanced class distributions, and exploring more advanced architectures (e.g., transfer learning backbones) and color/texture normalization methods to further improve robustness.

4. CONCLUSION

This study developed a Convolutional Neural Network (CNN) model for automated soil type classification using soil-surface images collected from Kaggle and web extraction. The best model achieved 91.61% test accuracy across six soil classes, demonstrating that CNN can capture relevant soil visual patterns for practical multi-class classification.

Preprocessing and data augmentation played a crucial role in improving performance, increasing accuracy from 46% (without preprocessing) to 91.61% (with preprocessing and augmentation). However, limitations remain due to dataset size, potential domain shift between Kaggle and web images, and visual similarity among certain soil classes. Future research should increase data volume and diversity per class, evaluate transfer learning architectures, and report additional metrics (macro-F1 and confusion matrix) to strengthen reliability for real-world deployment.

REFERENCES

[1] E. K. Gyasi and S. Purushotham, "Soil-MobiNet: A Convolutional Neural Network Model Base Soil Classification to Determine Soil Morphology and Its Geospatial Location," *Sensors*, vol. 23, no. 15, p. 6709, 2023, doi: 10.3390/s23156709.

[2] A. D. Ronaldo, Hamzah, and M. Diqi, "Effective Soil Type Classification Using Convolutional Neural Network," *International Journal of Informatics and Computation (IJICOM)*, vol. 3, no. 1, pp. 20–29, 2021, doi: 10.35842/ijicom.v3i1.33.

[3] G. D. Chate and S. S. Bhamare, "Classification of Soil Images Using Convolutional Neural Network," *International Journal of Image, Graphics and Signal Processing*, vol. 17, no. 5, pp. 26–41, 2025, doi: 10.5815/ijigsp.2025.05.03.

[4] Y. A. Gerhana, A. R. Atmadja, W. B. Zulfikar, and N. Ashanti, "The implementation of K-nearest neighbor algorithm in case-based reasoning model for forming automatic answer identity and searching answer similarity of algorithm case," in *2017 5th International Conference on Cyber and IT Service Management (CITSM)*, IEEE, Aug. 2017, pp. 1–5. doi: 10.1109/CITSM.2017.8089233.

[5] I. Erliana, A. Karim, and Z. Zainabun, "Klasifikasi Tanah Kebun Kopi Arabika di Kabupaten Gayo Lues Berdasarkan Sistem Klasifikasi Soil Taxonomy USDA," *Jurnal Ilmiah Mahasiswa Pertanian*, vol. 7, no. 1, pp. 696–703, Feb. 2022, doi: 10.17969/JIMFP.V7I1.19049.

[6] D. Avianto, "Implementasi Ekstraksi Ciri Histogram dan K-Nearest Neighbor untuk Klasifikasi Jenis Tanah di Kota Banjar, Jawa Barat," *Jurnal Buana Informatika*, vol. 10, no. 2, pp. 85–98, Oct. 2019, doi: 10.24002/JBI.V10I2.2141.

[7] Y. Yohannes, S. Devella, and A. H. Pandrean, "Penerapan Speeded-Up Robust Feature pada Random Forest Untuk Klasifikasi Motif Songket Palembang," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 5, no. 3, Jan. 2019, doi: 10.28932/JUTISI.V5I3.1978.

[8] M. P. Behera, A. Sarangi, D. Mishra, and S. K. Sarangi, "A Hybrid Machine Learning algorithm for Heart and Liver Disease Prediction Using Modified Particle Swarm Optimization with Support Vector Machine," *Procedia Comput. Sci.*, vol. 218, pp. 818–827, Jan. 2023, doi: 10.1016/J.PROCS.2023.01.062.

[9] Y. Nugroho Paseneke, A. Nugroho, F. Teknologi Informasi, and U. Kristen Satya Wacana Jl, "Pemetaan dan Klasifikasi Kesesuaian Jenis Tanah Terhadap Tanaman Menggunakan Metode Naïve Bayes di Desa Cukilan," *AITI*, vol. 19, no. 2, pp. 199–212, Nov. 2022, doi: 10.24246/aiti.v19i2.199-212.

[10] D. S. Jodas, L. A. Passos, A. Adeel, and J. P. Papa, "PL-kNN: A Python-based implementation of a parameterless k-Nearest Neighbors classifier," *Software Impacts*, vol. 15, p. 100459, Mar. 2023, doi: 10.1016/J.SIMPA.2022.100459.

[11] M. A. Jabbar, F. A. Falcata, R. S. Fuadi, A. T. Hidayatuloh, M. T. Laut, and P. Banten, "Analysis of the Level of Satisfaction of Using the Majalengka Digital Samsat Website Using the K-Nearest Neighbor Algorithm," *Journal of Multimedia Technology and Applied Software*, vol. 2, no. 1, pp. 46–52, Apr. 2025, doi: 10.71266/JMTAS.V2I1.46.

[12] A. Herdiman, D. S. A. Maylawati, D. R. Ramdania, W. B. Zulfikar, M. I. Al-Amin, and M. A. Ramdhani, "Household Waste Classification with Convolutional Neural Networks (CNN)," *2024 9th International Conference on Informatics and Computing, ICIC 2024*, 2024, doi: 10.1109/ICIC64337.2024.10956537.

[13] Y. P. Astuti, E. R. Subhiyakto, I. Wardatunizza, and E. Kartikadarma, "Implementasi Algoritma Convolutional Neural Network (CNN) Untuk Klasifikasi Jenis Tanah Berbasis Android," *Jurnal Informatika: Jurnal Pengembangan IT*, vol. 8, no. 3, pp. 220–225, Sep. 2023.

[14] M. El Sakka, M. Ivanovici, L. Chaari, and J. Mothe, "A Review of CNN Applications in Smart Agriculture Using Multimodal Data," *Sensors*, vol. 25, no. 2, p. 472, 2025, doi: 10.3390/s25020472.

[15] A. Lab and F. Sheth, "Comprehensive Soil Classification Datasets." Accessed: Jan. 21, 2025. [Online]. Available: https://www.kaggle.com/datasets/ai4a-lab/comprehensive-soil-classification-datasets