Vol. 10, No. 2, Oktober 2025, Page. 518-533
https://journal.uin-alauddin.ac.id/index.php/instek
p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

Model Efisien Berbasis *Mobile Vision Transformer* (MobileViT) untuk Klasifikasi Jenis Tanah dari Citra

Darmatasia[*1], M. Hasrul. H[2]

[1,2]Teknik Informatika, Universitas Islam Negeri Alauddin Makassar, Indonesia
Email: [1]darmatasia@uin-alauddin.ac.id, [2]muhammad.hasrul@uin-alauddin.ac.id

Abstrak

Identifikasi jenis tanah berperan krusial dalam sektor pertanian. Namun, metode klasifikasi konvensional seperti uji laboratorium dan observasi langsung masih memiliki keterbatasan dalam hal efisiensi waktu, biaya, dan skala. Penelitian ini mengusulkan model efisien untuk klasifikasi citra tanah dengan arsitektur *Mobile Vision Transformer*. Pendekatan *transfer learning* digunakan dalam membangun model untuk mengatasi keterbatasan jumlah data latih yang selanjutnya disesuaikan dengan data jenis tanah melalui *fine-tuning*. Dataset yang digunakan dalam penelitian ini diperoleh dari platform Kaggle, yang terdiri dari enam kelas yaitu *Arid Soil*, *Black Soil*, *Laterite Soil*, *Mountain Soil*, *Red Soil*, dan *Yellow Soil*. Dataset dibagi menjadi 80% untuk pelatihan dan 20% untuk pengujian, dengan evaluasi kinerja berdasarkan akurasi, presisi, *recall*, F1-score, dan waktu inferensi. Eksperimen dilakukan dengan membandingkan performa MobileViT terhadap model konvensional seperti CNN ringan (MobileNet) dan *Vision Transformer* standar. Hasil penelitian menunjukkan bahwa model yang diusulkan mampu mencapai keseimbangan antara akurasi dan efisiensi komputasi, dengan tingkat akurasi sebesar 97%. MobileViT menunjukkan efisiensi waktu inferensi yang lebih baik dibandingkan Vision Transformer standar, dengan kecepatan sekitar 5 kali lebih cepat. Kecepatan inferensi MobileViT mendukung penerapannya pada aplikasi *real-time* berbasis perangkat dengan daya komputasi terbatas tanpa menurunkan akurasi dengan signifikan. Meskipun demikian, penelitian ini memiliki keterbatasan pada ukuran dataset yang terbatas.

Kata kunci: Jenis Tanah, *Mobile Vision Transformer*, Model Efisien

Abstract

Soil type identification plays a crucial role in the agricultural sector. However, conventional classification methods such as laboratory testing and direct observation still face limitations in terms of time efficiency, cost, and scalability. This quantitative study proposes an efficient soil image classification model based on the Mobile Vision Transformer architecture, offering high accuracy with low computational cost. A transfer learning approach is employed to address the limited amount of training data, followed by fine-tuning to adapt the model to the soil type dataset. The dataset used in this study is sourced from the Kaggle platform and consists of six classes: Arid Soil, Black Soil, Laterite Soil, Mountain Soil, Red Soil, and Yellow Soil. The dataset is split into 80% for training and 20% for testing, with performance evaluated based on accuracy, precision, recall, F1-score, and inference time. Experiments compare the performance of MobileViT against conventional models such as a lightweight CNN (MobileNet) and a standard Vision Transformer. The results demonstrate that the proposed model achieves a balance between accuracy and computational efficiency, reaching an accuracy level of 97%. MobileViT exhibits significantly better inference efficiency compared to the standard Vision Transformer, being approximately 5 times faster. This fast inference capability supports the deployment of MobileViT in real-time applications on devices with limited computational resources, without compromising accuracy.

Keywords: *Soil Type, Mobile Vision Transformer, Efficient Model*

1. PENDAHULUAN

Tanah merupakan salah satu sumber daya alam yang berperan sangat penting dalam mendukung kehidupan, terutama dalam sektor pertanian, kehutanan, perencanaan tata ruang, serta pengelolaan sumber daya lingkungan. Setiap jenis tanah memiliki karakteristik fisik dan kimia yang berbeda, yang secara langsung memengaruhi kemampuan lahan dalam mendukung pertumbuhan tanaman atau pembangunan infrastruktur. Oleh karena itu, kemampuan untuk mengklasifikasikan jenis tanah dengan

akurat dan efisien menjadi sangat krusial dalam mendukung berbagai pengambilan keputusan berbasis data.

Metode klasifikasi tanah yang selama ini banyak digunakan di lapangan masih sangat bergantung pada teknik konvensional, seperti pengamatan langsung, uji laboratorium, atau pemetaan manual berdasarkan sampel tanah. Meskipun metode tersebut telah teruji dari sisi akurasi, namun prosesnya memerlukan waktu yang lama, biaya yang besar, serta membutuhkan keahlian khusus dalam interpretasi hasil [1], [2], dan [3]. Hal ini menyulitkan proses klasifikasi tanah dalam skala luas, terutama di daerah-daerah terpencil dengan keterbatasan akses dan tenaga ahli.

Seiring dengan perkembangan teknologi digital dan kecerdasan buatan (AI), pendekatan berbasis citra dan *machine learning* mulai digunakan untuk menggantikan sebagian besar proses manual tersebut. Gambar tanah yang diambil melalui kamera digital atau sensor optik dapat diolah menggunakan teknik *computer vision* untuk mengekstraksi ciri-ciri seperti tekstur, warna, dan pola permukaan. Kombinasi citra dan algoritma *machine learning* atau *deep learning* memungkinkan sistem untuk belajar langsung dari data dan melakukan klasifikasi tanah secara otomatis dan cepat.

Penelitian oleh [4] menunjukkan bahwa metode gabungan *machine learning* dan *deep learning* mampu meningkatkan akurasi klasifikasi jenis tanah berbasis citra. Peneliti mengusulkan model *Multi-Stacking Ensemble* dan algoritma seleksi fitur Q-HOG, serta membandingkan performa berbagai arsitektur seperti RNN, LSTM, GRU, VGG16, dan algoritma klasik seperti Naïve Bayes, KNN, dan SVM. Dataset citra tanah dikumpulkan dari wilayah Vriddhachalam. Hasilnya, model usulan mencapai akurasi 98,96%.

Penelitian oleh [5] menggunakan 805 sampel tanah dari proyek metro Gayrettepe–Bandara Istanbul yang menerapkan KNN imputer untuk mengatasi *missing value* dan SMOTE untuk mengatasi ketidakseimbangan kelas. Peneliti membandingkan algoritma XGBoost, LightGBM, dan CatBoost, dengan *cross validation*. Hasil penelitian menunjukkan bahwa akurasi klasifikasi terbaik melebihi 90% yang membuktikan bahwa model *gradient boosting* mampu meningkatkan efisiensi dan keakuratan klasifikasi tanah secara signifikan.

Penelitian yang dilakukan oleh [6] mengusulkan metode klasifikasi tanah menggunakan tiga pendekatan machine learning, yaitu *Support Vector Classification* (SVC), *Multilayer Perceptron* (MLP), dan *Random Forest* (RF). Dataset yang digunakan terdiri dari 4.888 sampel tanah dari proyek di Vietnam, dengan 15 parameter sifat tanah sebagai input untuk mengklasifikasikan lima jenis tanah: *lean clay* (CL), *elastic silt* (MH), *fat clay* (CH), *clayey sand* (SC), dan *silt* (ML). Evaluasi dilakukan menggunakan kurva pembelajaran, *confusion matrix*, dan metrik statistik seperti akurasi, presisi, *recall*, dan F1-*score*. Hasil menunjukkan bahwa ketiga model memiliki performa tinggi (rata-rata akurasi 96,8%), dengan SVC memberikan hasil terbaik dengan akurasi 98,4%.

Metode tradisional untuk klasifikasi jenis tanah umumnya bersifat invasif, memakan waktu, dan mahal. Sebagai alternatif, pendekatan *non*-invasif seperti *Ground Penetrating Radar* (GPR) menawarkan solusi berbasis sifat elektromagnetik tanah. Studi yang dilakukan oleh [7] mengusulkan model *Deep Convolutional Neural Network* (CNN) untuk klasifikasi otomatis jenis tanah dari citra GPR B-Scan. Dataset sintetik dikembangkan menggunakan GPRMax dan digunakan untuk melatih serta memvalidasi model. Hasil menunjukkan bahwa model CNN mampu mengklasifikasikan 7 jenis tanah secara akurat, dengan tingkat akurasi mencapai 97% pada data uji .

Penelitian terkini menunjukkan bahwa *deep learning* memberikan kontribusi signifikan dalam klasifikasi jenis tanah yang sebelumnya bergantung pada metode konvensional yang memakan waktu. Penelitian oleh [8] menganalisis 12 publikasi terpilih dari 150 artikel ilmiah terkait penerapan *deep learning* dalam klasifikasi tanah. Teknik seperti *lightweight models* [1] , *multi-task learning* [9], dan *transfer learning*[10] menjadi tren baru dalam bidang tersebut. Meskipun hasil yang dicapai menunjukkan akurasi tinggi dalam identifikasi kelas dan prediksi sifat tanah, tantangan seperti ketidakseimbangan data, keterbatasan jumlah citra, kurangnya interpretabilitas model, dan tingginya waktu komputasi masih menjadi isu utama.

Penggunaan citra berbasis digital memiliki banyak keterbatasan, seperti variasi visual yang tinggi akibat perbedaan pencahayaan, sudut pandang, dan kondisi lingkungan yang menyulitkan model dalam membedakan kelas tanah secara konsisten [11]. Tantangan lain mencakup keterbatasan data terlabel yang representatif, serta tingginya ketergantungan pada lokasi geografis, sehingga model seringkali gagal melakukan generalisasi dengan baik.

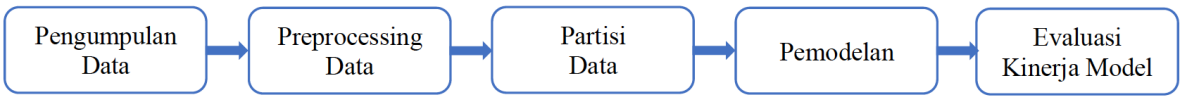
Salah satu pendekatan yang telah banyak digunakan dalam klasifikasi tanah adalah *Convolutional Neural Network* (CNN), yang terbukti mampu mengenali fitur visual dalam citra tanah. Namun, arsitektur CNN memiliki keterbatasan dalam memahami hubungan global antar area gambar karena pendekatannya yang fokus pada fitur lokal [12]. Untuk mengatasi keterbatasan ini, arsitektur baru bernama *Vision Transformer* (ViT) diperkenalkan dengan memanfaatkan prinsip *self-attention* dari arsitektur Transformer yang sebelumnya sukses di bidang *Natural Language Processing* (NLP). ViT mampu mengamati keseluruhan citra secara simultan, sehingga lebih efektif dalam memahami konteks spasial secara menyeluruh. Meskipun demikian, penerapan model berpresisi tinggi seperti Vision Transformer memerlukan daya komputasi besar [13] yang belum tentu tersedia di perangkat lapangan sehingga diperlukan arsitektur yang lebih efisien. Oleh karena itu, *Mobile Vision Transformer* (MobileViT) dikembangkan sebagai solusi efisien yang menggabungkan kekuatan CNN dan Transformer dalam arsitektur ringan, sehingga memungkinkan penerapan langsung di perangkat bergerak tanpa mengorbankan akurasi klasifikasi.

Dalam praktiknya, pelatihan model berbasis ViT termasuk MobileViT membutuhkan dataset berukuran sangat besar agar performa optimal dapat dicapai. Namun dalam konteks pengenalan jenis tanah, data citra yang tersedia seringkali terbatas sehingga dibutuhkan pendekatan khusus untuk mengatasi hal tersebut. *Transfer learning* adalah salah satu pendekatan dalam pembelajaran mesin yang memanfaatkan pengetahuan dari model yang telah dilatih sebelumnya pada tugas atau domain yang berbeda untuk diterapkan pada tugas baru yang serupa. Dengan memanfaatkan model yang telah dilatih sebelumnya pada dataset berskala besar seperti ImageNet, proses pelatihan menjadi lebih efisien karena fitur-fitur visual umum seperti bentuk, tekstur, dan pola warna sudah dipelajari oleh model dasar. Hal ini tidak hanya mempercepat waktu pelatihan, tetapi juga meningkatkan akurasi model pada dataset spesifik melalui transfer representasi visual yang kuat. Selain itu, transfer *learning* mengurangi kebutuhan sumber daya komputasi yang besar dan membantu mencegah *overfitting* yang sering terjadi pada dataset kecil. Dalam konteks *Vision Transformer* (ViT), pendekatan ini memungkinkan pemanfaatan arsitektur transformer yang kompleks tanpa harus melatih dari awal. Dengan demikian, transfer *learning* sangat cocok untuk tugas klasifikasi jenis tanah berbasis citra digital yang membutuhkan efisiensi, ketepatan, dan kemampuan generalisasi yang tinggi.

Penelitian ini memiliki dua kontribusi utama. Pertama, penelitian ini berkontribusi pada pengembangan metode klasifikasi jenis tanah berbasis kecerdasan buatan, khususnya melalui pemanfaatan citra digital dan model *deep learning* untuk meningkatkan akurasi identifikasi jenis tanah secara otomatis. Hal ini bertujuan untuk menggantikan atau melengkapi metode konvensional yang membutuhkan proses laboratorium yang mahal dan memakan waktu. Kedua, penelitian ini menghadirkan solusi model yang efisien dari segi waktu komputasi dengan mengadopsi arsitektur *Mobile Vision Transformer* (MobileViT) yang dirancang untuk berjalan secara optimal pada perangkat dengan keterbatasan sumber daya komputasi seperti perangkat *mobile*. Efisiensi waktu komputasi menjadi aspek yang sangat penting karena klasifikasi jenis tanah seringkali dibutuhkan secara cepat dan langsung di lapangan, baik untuk keperluan pemetaan cepat, pengambilan keputusan pertanian presisi, maupun *monitoring* lahan dalam skala luas. Selain itu, model yang ringan dan cepat dapat menghemat konsumsi energi, meningkatkan skalabilitas untuk pemrosesan data dalam jumlah besar, serta memungkinkan integrasi ke dalam aplikasi *mobile*, drone, atau perangkat *edge computing* secara *real-time*. Hal ini tidak hanya mempercepat proses pengambilan keputusan, tetapi juga menurunkan biaya operasional dan membuka peluang pemanfaatan teknologi klasifikasi tanah secara lebih luas, termasuk oleh petani kecil di daerah dengan keterbatasan infrastruktur.

2. METODE PENELITIAN

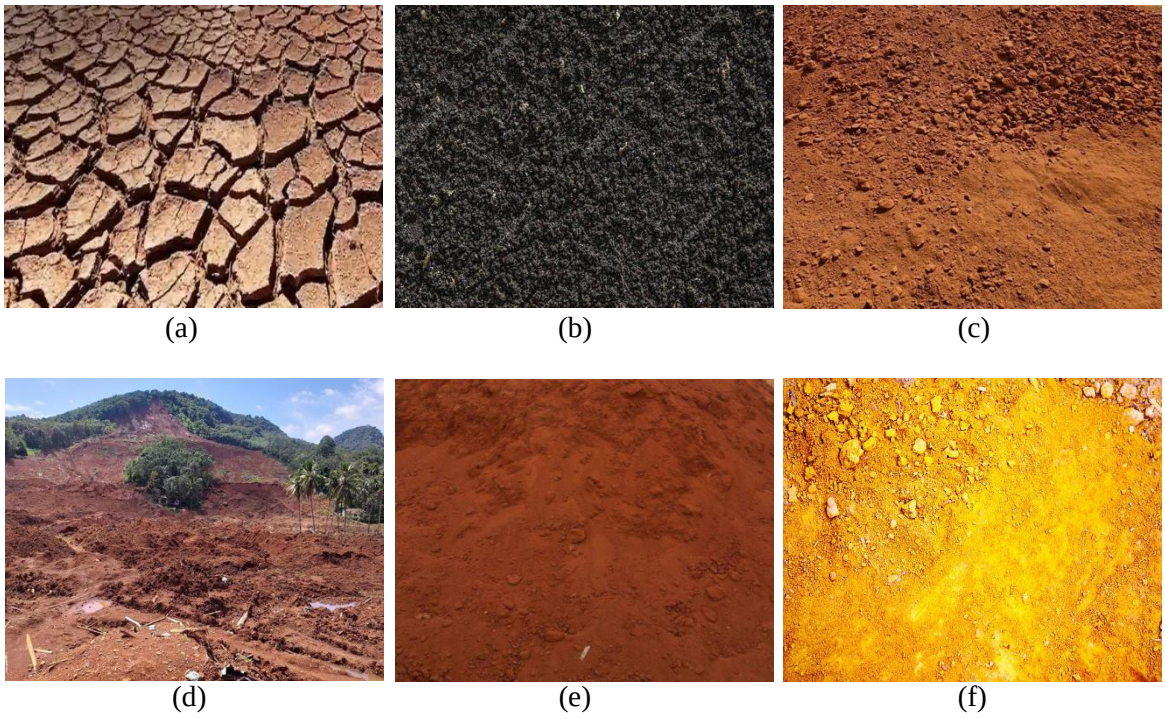
Penelitian ini terdiri dari enam tahapan utama yang dirancang secara sistematis untuk membangun model klasifikasi jenis tanah berbasis citra digital secara efisien dan akurat seperti yang dapat dilihat pada Gambar 1.



Gambar 1. Tahapan Penelitian

2.1 Pengumpulan Data

Tahap ini merupakan langkah awal dalam penelitian, yaitu mengumpulkan citra tanah yang akan digunakan sebagai data latih dan data uji. Penelitian ini menggunakan dataset citra jenis tanah yang diperoleh dari platform Kaggle [14] yang dibuat oleh [15]. Karena keterbatasan sumber daya perangkat yang digunakan untuk membangun model, maka data yang digunakan dalam penelitian ini hanya terdiri dari enam kelas jenis tanah yaitu *Arid Soil* (tanah kering), *Black Soil* (tanah hitam), *Laterite Soil* (tanah laterit), *Mountain Soil* (tanah pegunungan), *Red Soil* (tanah merah), dan *Yellow Soil* (tanah kuning), masing-masing berjumlah 50 citra. Contoh dataset dapat dilihat pada Gambar 2.



Gambar 2. Contoh Citra Jenis Tanah. (a) *Arid Soil* (tanah kering); (b) *Black Soil* (tanah hitam); (c) *Laterite Soil* (tanah laterit); (d) *Mountain Soil* (tanah pegunungan); (e) *Red Soil* (tanah merah); dan (f) *Yellow Soil* (tanah kuning).

2.2 *Preprocessing* Data

Tahap ini mencakup persiapan citra agar bisa digunakan oleh model *Mobile Vision Transformer*. *Preprocessing* mencakup *resizing* seluruh gambar ke ukuran 224x224 piksel agar sesuai dengan kebutuhan *input* model. Proses ini juga disertai normalisasi nilai piksel ke rentang [0, 1], yang merupakan standar dalam pengolahan citra untuk *deep learning*.

2.3 Partisi Data

Dataset dibagi menjadi dua bagian utama, yaitu data pelatihan (*training set*) dan data pengujian (*testing set*). Data pelatihan digunakan untuk melatih model agar mampu mempelajari pola-pola visual dari berbagai jenis tanah, sedangkan data pengujian digunakan untuk mengevaluasi kinerja model terhadap data yang belum pernah dilihat sebelumnya. Pemisahan dilakukan secara proporsional, yaitu 80% untuk pelatihan dan 20% untuk pengujian, sehingga dalam penelitian ini jumlah data per kelas terdiri atas 40 citra untuk pelatihan dan 10 citra untuk pengujian. Pemisahan dilakukan secara acak untuk menjaga distribusi kelas tetap seimbang pada masing-masing subset data.

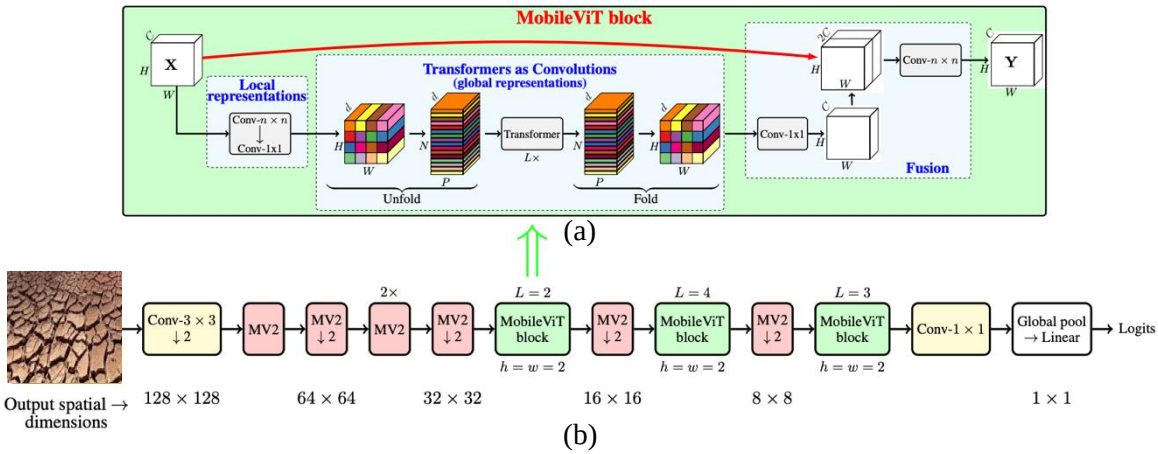
Selain itu, mengingat jumlah citra pada setiap kelas relatif terbatas, penelitian ini juga menerapkan 5-*fold cross-validation* untuk memperoleh evaluasi model yang lebih reliabel. Pada metode ini, seluruh dataset dibagi menjadi 5 *fold* dengan proporsi yang seimbang. Secara bergantian, empat *fold* digunakan

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

untuk pelatihan dan satu *fold* digunakan untuk pengujian. Pendekatan ini memastikan bahwa setiap citra berperan sebagai data uji tepat satu kali dan sebagai data latih sebanyak empat kali, sehingga kinerja model tidak hanya bergantung pada satu kali pembagian data. Dengan demikian, penggunaan 5-*fold cross-validation* membantu meminimalkan bias pemisahan data dan memberikan estimasi performa model yang lebih stabil pada kondisi data yang terbatas.

2.4 Pemodelan

Pada tahap ini, arsitektur MobileViT diimplementasikan. Model ini menggabungkan kekuatan *Convolutional Neural Networks* (CNN) untuk lokalitas dan Transformer untuk pemahaman spasial global dengan efisiensi tinggi.



Gambar 3. (a) *Vision Transformer* Standar; (b) *Mobile Vision Transformer* [16]

Vision Transformer (ViT) adalah adaptasi dari arsitektur *Transformer*, yang awalnya populer di bidang *Natural Language Processing* (NLP), untuk digunakan dalam tugas pengolahan citra. ViT diperkenalkan oleh Alexey Dosovitskiy [17]. *Transformer* bekerja dengan sangat baik pada data berbentuk urutan (*sequence*) seperti teks. ViT mengadopsi pendekatan serupa dengan mengubah gambar menjadi urutan *patch* (potongan kecil dari gambar) yang kemudian diproses seperti token dalam NLP.

Arsitektur *Vision Transformer* (ViT) bekerja dengan membagi gambar menjadi potongan kecil (*patch*), kemudian meratakan tiap *patch* menjadi vektor. Vektor-vektor tersebut diproyeksikan ke dalam bentuk *embedding* berdimensi rendah dan diberi informasi posisi (*positional embedding*). Seluruh urutan *embedding* dimasukkan ke dalam *encoder Transformer* standar.

Model ViT standar, yang ditunjukkan pada Gambar 3(a), mengubah input $X \in R^{H \times W \times C}$ menjadi urutan *patch* yang telah diratakan $X_f \in R^{N \times PC}$, kemudian memproyeksikannya ke dalam ruang berdimensi tetap d menjadi $X_p \in R^{N \times d}$, dan selanjutnya mempelajari representasi antar-*patch* menggunakan susunan sebanyak L blok transformer. Biaya komputasi dari *self-attention* pada Vision Transformer adalah $O(N^2 d)$. C, H, dan W masing-masing merepresentasikan jumlah kanal, tinggi, dan lebar tensor. $P=wh$ adalah jumlah piksel dalam satu *patch* dengan tinggi h dan lebar w serta N adalah jumlah total *patch*.

Penelitian oleh [16] memperkenalkan model ViT ringan, yaitu MobileViT. Gagasan utamanya adalah mempelajari representasi global menggunakan *transformer* yang bekerja seperti konvolusi. Pendekatan tersebut memungkinkan untuk mengintegrasikan sifat-sifat seperti konvolusi (misalnya bias spasial) secara implisit ke dalam jaringan, mempelajari representasi dengan pelatihan yang sederhana (misalnya augmentasi dasar), dan dengan mudah mengintegrasikan MobileViT ke dalam arsitektur lanjutan lainnya (misalnya DeepLabv3 untuk tugas segmentasi).

Blok MobileViT seperti yang ditunjukkan pada Gambar 3(b) dirancang untuk memodelkan informasi lokal dan global dari sebuah tensor masukan dengan jumlah parameter yang lebih sedikit. Secara formal, untuk sebuah tensor masukan $X \in R^{H \times W \times C}$, MobileViT menerapkan sebuah lapisan

konvolusi standar berukuran n×n yang kemudian diikuti oleh lapisan konvolusi *point-wise* (atau 1×1) untuk menghasilkan tensor $X_L \in R^{H \times W \times d}$. Lapisan konvolusi n×n berfungsi untuk menangkap informasi spasial lokal, sementara konvolusi *point-wise* memproyeksikan tensor ke dalam ruang berdimensi tinggi (yakni berdimensi-d, di mana d>C) dengan mempelajari kombinasi linear dari kanal-kanal masukan**.** MobileViT bertujuan untuk memodelkan dependensi jarak jauh yang bersifat *non*-lokal sambil tetap memiliki luas bidang reseptif yang efektif sebesar H × W [16].

Konvolusi standar dapat dipandang sebagai rangkaian tiga operasi berurutan, yaitu: (1) *unfolding*; (2) perkalian matriks; (3) *folding*. Blok MobileViT memiliki kemiripan dengan konvolusi karena sama-sama memanfaatkan komponen dasar yang serupa. Namun, blok MobileViT menggantikan pemrosesan lokal (perkalian matriks) dalam konvolusi dengan pemrosesan global yang lebih dalam (melalui tumpukan lapisan transformer). Sebagai hasilnya, MobileViT memiliki sifat-sifat seperti konvolusi, misalnya bias spasial [16]. Oleh karena itu, blok MobileViT dapat dipandang sebagai transformer yang berperilaku seperti konvolusi. Salah satu keunggulan dari desain yang sengaja dibuat sederhana tersebut adalah bahwa implementasi efisien pada tingkat rendah dari konvolusi dan transformer dapat langsung digunakan. Hal ini memungkinkan penggunaan MobileViT di berbagai perangkat.

Biaya komputasi dari *self-attention multi-head* pada MobileViT dan ViT masing-masing adalah $O(N^2Pd)$ dan $O(N^2d)$. Secara teoretis, MobileViT tampak kurang efisien dibandingkan ViT. Namun, dalam praktiknya, MobileViT justru lebih efisien daripada ViT. MobileViT memiliki dua kali lebih sedikit FLOPs dan memberikan akurasi 1,8% lebih baik dibandingkan DeIT pada dataset ImageNet-1K [16] dan [18].

Arsitektur *Mobile Vision Transformer* yang digunakan dalam penelitian adalah *MobileViT-small*, *optimizer* yang digunakan adalah Adam dengan *learning rate* sebesar 2e-5 (0.00002). Proses pelatihan dilakukan selama 5 dan 10 *epoch*. Ukuran *batch* yang digunakan adalah 32, artinya setiap iterasi pelatihan memproses 32 gambar sekaligus sebelum memperbarui bobot model. Konfigurasi ini disesuaikan agar seimbang antara kecepatan pelatihan dan kapasitas memori, sekaligus cukup efektif untuk menyerap fitur visual dari gambar tanah yang bervariasi secara tekstur dan warna.

Selain model *Mobile Vision Transformer*, dalam penelitian ini juga dibandingkan dengan arsitektur *Vision Transformer* standar dan CNN yaitu MobileNet dengan menggunakan pengaturan parameter yang sama.

2.5 Evaluasi Kinerja Model

Metode evaluasi yang digunakan dalam penelitian ini bertujuan untuk mengukur kinerja model dalam mengklasifikasikan jenis tanah berdasarkan citra. Evaluasi dilakukan menggunakan metrik akurasi, presisi, *recall*, F1-*score*, *confusion matrix*, dan waktu inferensi

Akurasi digunakan untuk mengetahui seberapa besar proporsi prediksi yang benar terhadap seluruh data uji. Namun karena akurasi saja tidak selalu cukup, terutama ketika jumlah sampel per kelas tidak seimbang, maka digunakan pula presisi dan *recall*. Presisi mengukur seberapa banyak prediksi positif yang benar (ketepatan klasifikasi tiap kelas), sedangkan *recall* menunjukkan seberapa banyak sampel aktual yang berhasil dikenali oleh model [19].

Untuk memberikan gambaran yang lebih menyeluruh, dihitung pula F1-*score*, yaitu rata-rata harmonis antara presisi dan *recall*, sehingga sangat berguna dalam menilai keseimbangan performa model pada semua kelas.

Selain metrik numerik, digunakan juga *confusion matrix* untuk melihat pola kesalahan klasifikasi antar kelas. *Confusion matrix* menggambarkan secara visual jumlah prediksi yang benar dan salah untuk masing-masing kelas, sehingga dapat diidentifikasi apakah terdapat kelas tertentu yang sering tertukar dengan kelas lainnya.

Dengan kombinasi metrik ini, evaluasi model menjadi lebih komprehensif dan dapat digunakan untuk menilai efektivitas model dalam mengenali berbagai jenis tanah secara akurat. Selanjutnya, waktu inferensi digunakan untuk mengukur waktu yang dibutuhkan oleh model untuk melakukan prediksi terhadap satu *input* setelah proses pelatihan selesai.

Vol. 10, No. 2, Oktober 2025, Page. 518-533
https://journal.uin-alauddin.ac.id/index.php/instek
p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

3. HASIL DAN PEMBAHASAN

Hasil pelatihan model *MobileViT* selama 5 *epoch* dapat dilihat pada Tabel 1. Model menunjukkan peningkatan akurasi pada setiap *epoch*. Pada *epoch* pertama, model sudah mencapai akurasi sebesar 81.25%, yang diikuti dengan peningkatan akurasi hingga mencapai 95.83% pada *epoch* terakhir.

Tabel 1. Hasil *Training* Model *MobileViT* dengan 5 *epoch*

Epoch	Akurasi
1	0.8125
2	0.8708
3	0.9042
4	0.9375
5	**0.9583**

Berdasarkan Tabel 1, dapat dilihat bahwa Akurasi meningkat secara konsisten dari *epoch* ke *epoch*, menunjukkan bahwa model mampu belajar dengan baik dari data pelatihan. Lonjakan terbesar terjadi antara *epoch* 1 ke 2 (+5.83%) dan 2 ke 3 (+3.34%), yang mengindikasikan bahwa pada awal pelatihan, model menangkap pola dasar dengan cepat. Setelah *epoch* ke-3, peningkatan mulai melambat, yang umum terjadi karena model mulai mendekati konvergensi. Pada *epoch* ke-5, akurasi sudah mencapai 95.83%, menunjukkan performa pelatihan yang sangat baik untuk klasifikasi jenis tanah berbasis citra. Model *MobileViT* menunjukkan kemampuan pembelajaran yang stabil dan efisien selama 5 *epoch* pelatihan.

Hasil pelatihan *MobileViT* selama 10 *epoch* dapat dilihat pada Tabel 2. Model menunjukkan peningkatan akurasi pada setiap *epoch*. Pada *epoch* pertama, model mencapai akurasi sebesar 97.50%, yang diikuti dengan peningkatan akurasi hingga mencapai 100% pada *epoch* terakhir.

Tabel 2. Hasil *Training* Model *MobileViT* dengan 10 *epoch*

Epoch	Akurasi
1	0.9750
2	0.9750
3	0.9833
4	0.9792
5	1.0000
6	0.9958
7	0.9917
8	0.9958
9	0.9875
10	**1.0000**

Berdasarkan Tabel 2 dapat dilihat bahwa Akurasi awal sudah tinggi sejak *epoch* ke-1 (97.50%), menunjukkan bahwa model MobileViT telah melakukan *transfer learning* dengan sangat baik. Stabilitas performa terlihat jelas karena akurasi tetap berada di atas 97% sepanjang pelatihan. Akurasi maksimal 100% tercapai pada *epoch* ke-5 dan kembali tercapai pada *epoch* ke-10 yang menunjukkan bahwa model telah mampu mengenali seluruh data pelatihan dengan benar pada dua titik tersebut. Meski terjadi sedikit fluktuasi ringan antar *epoch* (misalnya penurunan kecil di *epoch* 9 menjadi 98.75%), variasi tersebut masih dalam rentang yang sangat tinggi dan tidak mengindikasikan *overfitting* yang parah. Kinerja model cenderung mendekati konvergensi, karena tidak terjadi peningkatan signifikan setelah *epoch* ke-5. Model MobileViT menunjukkan performa pelatihan yang sangat baik dalam 10 *epoch*, dengan akurasi yang sangat tinggi dan relatif stabil. Pencapaian akurasi sempurna pada beberapa *epoch* mengindikasikan bahwa model berhasil mempelajari pola dari data pelatihan secara menyeluruh.

Hasil pelatihan model *Vision Transformer* standar selama 5 *epoch* dapat dilihat pada Tabel 3. Model menunjukkan peningkatan akurasi pada setiap *epoch*. Pada *epoch* pertama, model mencapai akurasi sebesar 23.75%, yang diikuti dengan peningkatan akurasi hingga mencapai 95.83% pada *epoch* terakhir.

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

Tabel 3. Hasil *Training* Model Vision Transformer standar dengan 5 *epoch*

Epoch	Akurasi
1	0.2375
2	0.7667
3	0.8875
4	0.9375
5	**0.9583**

Berdasarkan Tabel 3, dapat dilihat bahwa pada *epoch* 1 model masih dalam tahap awal belajar yang menandakan bahwa bobot model belum optimal dalam mengenali pola visual dari citra tanah. Pada *epoch* 2-3 terjadi lonjakan akurasi signifikan (dari 76.67% ke 88.75%), menunjukkan bahwa model mulai memahami fitur penting dari data. Epoch 4-5 akurasi meningkat secara stabil dan mendekati maksimum. Pada *epoch* ke-5, model mencapai akurasi 95.83%, menandakan model sudah cukup konvergen dan tidak terjadi *overfitting* pada tahap ini.

Hasil pelatihan *Vision Transformer* selama 10 *epoch* dapat dilihat pada Tabel 4. Model menunjukkan peningkatan akurasi pada setiap *epoch*. Pada *epoch* pertama, model mencapai akurasi sebesar 29.17%, yang diikuti dengan peningkatan akurasi hingga mencapai 99.58% pada *epoch* terakhir.

Tabel 4. Hasil *Training* Model Vision Transformer dengan 10 *epoch*

Epoch	Akurasi
1	0.2917
2	0.6417
3	0.8750
4	0.9208
5	0.9500
6	0.9708
7	0.9708
8	0.9875
9	0.9875
10	**0.9958**

Berdasarkan Tabel 4, dapat dilihat pada *epoch* 1-2 akurasi masih di bawah 65%, menunjukkan bahwa model baru mulai membangun pemahaman terhadap pola-pola visual pada data citra. *Epoch* 3-5 terjadi kenaikan signifikan, model dengan cepat belajar hingga akurasi melewati 95% pada *epoch* ke-5. *Epoch* 6-10 model menunjukkan stabilitas yang tinggi, dengan nilai akurasi mendekati sempurna (>97%) dan akhirnya mencapai 99.58% di *epoch* ke-10. Kenaikan dari *epoch* ke *epoch* tampak mulai melambat setelah *epoch* ke-6, menandakan model mendekati konvergensi.

Pelatihan hingga 10 *epoch* menghasilkan model dengan akurasi sangat tinggi dan kestabilan performa yang baik. Meskipun akurasi telah tinggi sejak *epoch* ke-5, pelatihan tambahan hingga *epoch* ke-10 tetap memberikan sedikit peningkatan, yang bisa berguna dalam konteks klasifikasi citra yang membutuhkan presisi tinggi. Namun, perlu dipertimbangkan *trade-off* antara waktu komputasi dan peningkatan akurasi yang marginal setelah model stabil.

Hasil pelatihan MobileNet selama 5 *epoch* dapat dilihat pada Tabel 5. Model menunjukkan peningkatan akurasi pada setiap *epoch*. Pada *epoch* pertama, model mencapai akurasi sebesar 27.49% yang diikuti dengan peningkatan akurasi hingga mencapai 51.67% pada *epoch* terakhir.

Tabel 5. Hasil *Training* Model MobileNet dengan 5 *epoch*

Epoch	Akurasi
1	0.2749
2	0.2500
3	0.4157
4	0.1875
5	**0.5167**

Berdasarkan Tabel 5, pada *epoch* 1-2 akurasi masih rendah (di bawah 30%), menunjukkan bahwa model baru mulai mengenali pola dan membutuhkan pelatihan lebih lanjut. Pada *epoch* 3 terdapat peningkatan signifikan (menjadi 0.4157), mengindikasikan bahwa model mulai belajar dengan lebih baik. Pada *epoch* 4 terjadi penurunan drastis (0.1875), yang bisa disebabkan oleh fluktuasi selama pelatihan, seperti *overfitting* sementara atau *learning rate* yang kurang optimal. Pada *epoch* 5 model menunjukkan pemulihan performa, mencapai akurasi tertinggi pada iterasi tersebut yaitu 51.67%.

Model MobileNet mengalami fluktuasi akurasi selama 5 *epoch*, dengan peningkatan yang belum stabil. Akurasi maksimum sebesar 51.67% masih tergolong sedang untuk tugas klasifikasi berbasis citra. Hal ini dapat disebabkan oleh beberapa faktor seperti jumlah *epoch* yang terlalu sedikit, parameter pelatihan yang belum optimal atau kompleksitas model yang belum cukup untuk mengekstraksi fitur dari dataset secara efisien.

Dibandingkan dengan Mobile Vision Transformer yang mencapai akurasi hampir sempurna, MobileNet membutuhkan *epoch* tambahan atau penyesuaian *hyperparameter* untuk meningkatkan performanya.

Hasil pelatihan MobileNet selama 10 *epoch* dapat dilihat pada Tabel 6. Model menunjukkan peningkatan akurasi pada setiap *epoch*. Pada *epoch* pertama, model mencapai akurasi sebesar 68.66% yang diikuti dengan peningkatan akurasi hingga mencapai 87.50% pada *epoch* terakhir.

Tabel 6. Hasil *Training* Model MobileNet dengan *epoch* 10

Epoch	**Akurasi**
1	0.6866
2	0.5625
3	0.7153
4	0.6875
5	0.7741
6	0.5625
7	0.7819
8	0.8125
9	0.8438
10	**0.8750**

Berdasarkan Tabel 6, *epoch* 1-2, model sudah menunjukkan akurasi cukup baik sejak awal (0.6866), namun menurun di *epoch* ke-2. *Epoch* 3-5 terjadi peningkatan kembali secara bertahap, menunjukkan model mulai belajar dengan lebih stabil. Pada *epoch* 6, akurasi turun ke 0.5625, mungkin disebabkan oleh variasi *batch* data atau *learning rate* yang belum sempurna. Pada *epoch* 7-10 terjadi peningkatan konsisten dan signifikan, hingga mencapai akurasi tertinggi 87.50% pada *epoch* ke-10.

Pelatihan model MobileNet hingga 10 *epoch* menunjukkan tren peningkatan akurasi secara signifikan dan konsisten, khususnya setelah *epoch* ke-6. Hal ini menunjukkan bahwa model belum mencapai performa optimal saat 5 *epoch* (Tabel 5), namun mulai menunjukkan konvergensi dan stabilitas pada *epoch* 7 ke atas, akurasi akhir (87.50%) menunjukkan bahwa MobileNet cukup mampu mengenali fitur data dengan baik setelah diberi waktu pelatihan yang cukup. Jumlah *epoch* yang lebih banyak memberikan hasil yang jauh lebih baik dibanding hanya 5 *epoch*.

Selain pengukuran akurasi, pada penelitian ini juga dilakukan pengukuran presisi, *recall*, dan F-1 *Score*, serta waktu inferensi pada ketiga model. Hasil evaluasi model *MobileViT* dapat dilihat pada Tabel 7, Hasil evaluasi Model Vision transformer pada Tabel 8, sedangkan hasil evaluasi *MobileNet* dapat dilihat pada Tabel 9.

Tabel 7. Hasil Evaluasi Model *MobileViT* (10 *epoch*)

Jenis Tanah	**Presisi**	**Recall**	**F1-Score**
Arid	1.00	0.80	0.89
Black	1.00	1.00	1.00
Laterite	1.00	1.00	1.00
Mountain	0.83	1.00	0.91
Red	1.00	0.90	1.00
Yellow	1.00	1.00	1.00

Berdasarkan Tabel 7, dapat dilihat bahwa model sangat akurat saat memprediksi kelas *Arid* (tidak ada *false positive*), namun tidak semua Arid terdeteksi (ada *false negative*). Pada kelas *Black*,*Laterite*, dan *Yellow*, semua prediksi benar dan semua terdeteksi dengan benar. Pada kelas *Mountain*, nilai *recall* tinggi, tapi ada citra diprediksi salah sebagai Mountain (presisi < 1). Model tidak pernah salah dalam memprediksi kelas *Red* (presisi 1.00), tetapi ada beberapa *Red* yang tidak dikenali.

Tabel 8. Hasil Evaluasi Model *Vision Transformer* Standar (10 *epoch*)

Jenis Tanah	Presisi	Recall	F1-Score
Arid	1.00	1.00	1.00
Black	1.00	1.00	1.00
Laterite	0.91	1.00	0.95
Mountain	1.00	1.00	1.00
Red	1.00	0.90	0.95
Yellow	1.00	1.00	1.00

Berdasarkan Tabel 8, Laterite memiliki presisi 0.91, artinya model terkadang keliru memprediksi sampel sebagai Laterite padahal bukan. Namun *recall*-nya 1.00, artinya semua tanah *Laterite* yang sebenarnya berhasil dikenali. *Red* memiliki *recall* yang lebih rendah (0.90), artinya sebagian sampel tanah *Red* tidak dikenali sebagai *Red*, meskipun model tidak pernah salah memprediksi selain Red sebagai Red (presisi 1.00).

Tabel 9. Hasil Evaluasi Model MobileNet (10 *epoch*)

Jenis Tanah	Presisi	Recall	F1-Score
Arid	**0.89**	0.80	**0.84**
Black	0.62	0.50	0.56
Laterite	0.80	0.80	0.80
Mountain	0.49	0.60	0.48
Red	0.75	**0.90**	0.82
Yellow	0.83	0.50	0.62

Jenis tanah *Laterite* dan *Red* memiliki nilai F1-*score* yang relatif baik yaitu masing-masing 0.80 dan 0.82. Tanah *Arid* juga cukup baik (F1-*score* = 0.84), meskipun *recall*-nya 0.80 (masih ada data Arid yang tidak dikenali). Jenis tanah Black, Mountain, dan Yellow memiliki F1-Score yang rendah (terendah adalah Mountain = 0.48), menandakan banyak kesalahan dalam klasifikasi. Jenis tanah Black rendah di semua metrik dengan nilai presisi 0.62 dan *recall* 0.50. Jenis tanah *Mountain* memiliki presisi sangat rendah (0.49) yang berarti banyak prediksi yang salah. Jenis tanah *Yellow* memiliki nilai *Recall* sangat rendah (0.50) yang berarti banyak sampel *Yellow* yang tidak dikenali.

MobileNet menunjukkan kinerja bervariasi antar kelas. Performa terbaik pada jenis tanah *Arid, Laterite, dan Red* tetapi masih lebih rendah dibandingkan dengan performa MobileViT dan *Vision Transformer* (ViT) standar pada semua kelas. Adapun nilai F1-*score* tertinggi adalah 0.84 yaitu pada jenis tanah *Arid* dan terendah yaitu 0.48 pada jenis tanah *Mountain*).

Kinerja MobileNet masih cenderung tidak stabil pada 10 *epoch*, dengan indikasi model kurang mampu membedakan kelas yang memiliki kemiripan visual.

Secara keseluruhan, perbandingan hasil evaluasi akurasi model MobileViT, *Vision Transformer* Standar, dan MobileNet pada *epoch* 5 dan 10 dapat dilihat pada Tabel 10.

Tabel 10. Perbandingan Hasil Akurasi Model MobileViT, *Vision Transformer* Standar, dan MobileNet

Model	***Epoch***	**Akurasi**
MobileViT	5	0.86
	10	**0.97**
Vision Transformer Standar	5	0.88
	10	**0.98**
MobileNet	5	0.57
	10	**0.68**

Vol. 10, No. 2, Oktober 2025, Page. 518-533
https://journal.uin-alauddin.ac.id/index.php/instek
p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

Tabel 10 menunjukkan perbandingan akurasi tiga model yaitu MobileViT, Vision Transformer Standar, dan MobileNet pada berbagai jumlah *epoch* pelatihan. Hasil evaluasi menunjukkan bahwa model Vision Transformer Standar menghasilkan akurasi tertinggi sebesar 0.98 setelah 10 *epoch*, diikuti oleh MobileViT dengan akurasi 0.97 pada jumlah *epoch* yang sama. Meskipun selisih akurasinya sangat kecil, MobileViT memiliki keunggulan dari sisi efisiensi karena dirancang sebagai versi ringan dari Vision Transformer. Kedua model ini juga menunjukkan akurasi awal yang tinggi pada *epoch* ke-5, yaitu 0.86 untuk MobileViT dan 0.88 untuk Vision Transformer Standar, yang menandakan kemampuan keduanya dalam memahami pola sejak tahap awal pelatihan.

Sementara itu, model MobileNet menunjukkan performa yang jauh lebih rendah. Pada *epoch* ke-5, akurasinya hanya mencapai 0.57, dan meskipun meningkat hingga 0.68 pada *epoch* ke-10, peningkatannya tergolong lambat dan tidak signifikan. Hal ini menunjukkan bahwa MobileNet kurang cocok digunakan untuk tugas klasifikasi jenis tanah pada dataset ini. Dengan demikian, dapat disimpulkan bahwa MobileViT merupakan pilihan yang optimal karena mampu memberikan akurasi tinggi yang sebanding dengan Vision Transformer Standar, namun dengan efisiensi pelatihan yang lebih baik.

Perbandingan waktu komputasi pelatihan model *MobileViT*, *Vision Transformer* Standar, dan MobileNet dapat dilihat pada Tabel 11.

Tabel 11. Perbandingan Waktu *Training* Model MobileViT, *Vision Transformer* Standar, dan MobileNet

Epoch	**Waktu Komputasi Pelatihan**		
	MobielViT	*Vision Transformer Standar*	MobileNet
1	139s	754s	28s
2	142s	621s	3s
3	146s	647s	27s
4	138s	601s	3s
5	144s	636s	13s
6	142s	604s	3s
7	146s	636s	17s
8	141s	640s	6s
9	138s	649s	15s
10	127s	643s	3s

Tabel 11 menyajikan perbandingan waktu komputasi pelatihan untuk tiga model yaitu MobileViT, Vision Transformer Standar, dan MobileNet selama 10 *epoch* pelatihan. Dari data yang ditampilkan, dapat diketahui bahwa Vision Transformer Standar memiliki waktu pelatihan paling lama, dengan rata-rata sekitar 643 detik per *epoch*, dan total waktu pelatihan mencapai sekitar 6.431 detik atau hampir 1 jam 47 menit. Hal ini mencerminkan kompleksitas arsitekturnya yang tinggi, meskipun diimbangi dengan akurasi terbaik seperti ditunjukkan pada tabel sebelumnya.

MobileViT menunjukkan efisiensi yang jauh lebih baik, dengan rata-rata waktu pelatihan sekitar 139 detik per *epoch*, dan total selama 10 *epoch* mencapai 1.393 detik atau sekitar 23 menit. Hal tersebut menunjukkan bahwa MobileViT merupakan pilihan yang seimbang antara kinerja dan efisiensi, karena waktu pelatihannya jauh lebih cepat dibanding Vision Transformer namun akurasinya tetap tinggi.

Di sisi lain, model MobileNet memiliki waktu pelatihan paling singkat, dengan rata-rata hanya sekitar 11,5 detik per *epoch*, dan total selama 10 *epoch* hanya 115 detik atau kurang dari 2 menit. Meskipun sangat efisien dari sisi komputasi, seperti telah diuraikan sebelumnya, performa akurasinya masih rendah, sehingga model ini lebih cocok digunakan pada perangkat dengan keterbatasan sumber daya yang tidak membutuhkan tingkat akurasi tinggi.

Perbandingan waktu inferensi model MobileViT, *Vision Transformer* standar, dan MobileNet dapat dilihat pada Tabel 12.

Tabel 12. Perbandingan Waktu inferensi Model MobileViT, *Vision Transformer* Standar, dan MobileNet

Model	***Waktu (ms)***
MobileViT	141ms
Vision Transformer Standar	783ms
MobileNet	67ms

Tabel 12 memperlihatkan perbandingan waktu inferensi dari tiga model yaitu MobileViT, Vision Transformer Standar, dan MobileNet. Waktu inferensi merupakan durasi yang dibutuhkan oleh model untuk melakukan prediksi terhadap satu *input* setelah proses pelatihan selesai. Dari data terlihat bahwa MobileNet memiliki waktu inferensi tercepat, yaitu hanya 67 milidetik, diikuti oleh MobileViT dengan waktu 141 milidetik, dan Vision Transformer Standar dengan waktu yang paling lambat yaitu 783 milidetik. Hal tersebut sejalan dengan penelitian yang dilakukan oleh [17] bahwa ViT memiliki kemampuan luar biasa dalam memahami struktur visual, tetapi membutuhkan komputasi yang besar dan waktu inferensi yang lama.

Meskipun MobileNet unggul dalam kecepatan inferensi, keunggulan ini perlu dilihat dalam konteks performa keseluruhan. Seperti yang telah diuraikan sebelumnya, akurasi MobileNet cukup rendah, sehingga kecepatan tidak serta merta menjadikannya pilihan terbaik. Hal tersebut sejalan dengan penelitian dari [20] bahwa MobileNet dirancang untuk efisiensi tinggi pada perangkat dengan keterbatasan sumber daya komputasi, sehingga sangat cepat dalam proses inferensi. Namun, efisiensi ini disertai dengan penurunan tingkat akurasi, terutama pada tugas klasifikasi yang kompleks. Artinya, kecepatan tinggi yang ditawarkan MobileNet tidak selalu sejalan dengan performa klasifikasinya dalam konteks yang membutuhkan ketepatan tinggi. Hal tersebut menjelaskan mengapa meskipun MobileNet menunjukkan waktu inferensi tercepat dalam Tabel 12, MobileNet kurang tepat digunakan dalam kasus yang memerlukan akurasi tinggi seperti klasifikasi jenis tanah. Sebaliknya, model seperti MobileViT menjadi alternatif yang lebih seimbang karena mampu mempertahankan akurasi tinggi dengan waktu inferensi yang tetap efisien. Vision Transformer memiliki akurasi sangat tinggi, tetapi waktu inferensinya jauh lebih lambat, yang dapat menjadi kendala dalam aplikasi *real-time* atau pada perangkat dengan keterbatasan daya komputasi.

Model MobileViT menawarkan titik tengah yang ideal, dengan waktu inferensi yang cukup cepat namun tetap mampu menghasilkan akurasi tinggi. Dengan demikian, MobileViT sangat cocok untuk diterapkan pada sistem yang memerlukan respon cepat namun tetap akurat, seperti pada perangkat *edge* atau aplikasi *mobile* berbasis AI.

Selain menggunakan pembagian data secara konvensional, penelitian ini juga menerapkan 5-*fold cross-validation* untuk memperoleh gambaran kinerja model yang lebih komprehensif dan reliabel. Hasil evaluasi yang diperoleh dapat dilihat pada Tabel 12.

Tabel 12. Perbandingan Akurasi Model MobileViT, *Vision Transformer* Standar, dan MobileNet dengan 5-*fold cross-validation*

Fold	**Mobile ViT (%)**	**Vision Transformer Standar (%)**	**MobileNet (%)**
Fold 1	93.33	98.44	61.66
Fold 2	100	100	73.33
Fold 3	100	100	61.66
Fold 4	100	100	66.66
Fold 5	100	100	68.33
Rata-rata	**98.67**	**99.68**	**66.33**

Berdasarkan Tabel 12, hasil evaluasi menunjukkan bahwa MobileViT dan Vision Transformer (ViT) memiliki akurasi yang jauh lebih tinggi dibandingkan MobileNet pada setiap fold, dengan rata-rata akurasi masing-masing sebesar 98.67% dan 99.68%, jauh di atas MobileNet yang hanya mencapai 66.33%. Kinerja unggul MobileViT dan ViT terutama disebabkan oleh kemampuannya dalam menangkap hubungan global antar-piksel pada citra tanah yang sering kali memiliki tekstur halus dan pola spasial yang tidak teratur. Arsitektur transformer memungkinkan kedua model tersebut mempelajari dependensi jangka panjang antarbagian gambar secara lebih efektif dibandingkan CNN

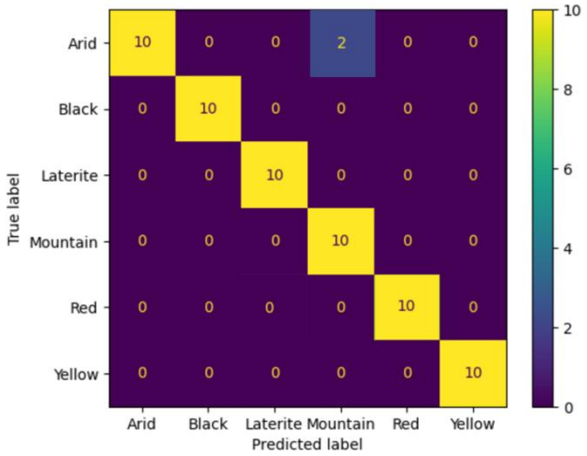
tradisional. MobileViT menggabungkan keunggulan CNN pada ekstraksi fitur lokal dengan kekuatan transformer dalam memahami konteks global, sehingga mampu menghasilkan representasi fitur yang lebih diskriminatif meskipun ukuran model lebih ringan. Sementara itu, ViT standar memiliki kapasitas lebih besar dan kemampuan representasi yang lebih mendalam yang menjelaskan mengapa akurasinya mendekati sempurna pada seluruh fold. Sebaliknya, MobileNet sebagai CNN ringan lebih terbatas dalam menangkap pola global dan cenderung lebih sensitif terhadap variasi tekstur tanah, sehingga menghasilkan performa yang lebih rendah.

Secara keseluruhan, dapat disimpulkan bahwa model berbasis Transformer memberikan performa lebih unggul dan konsisten untuk klasifikasi citra tanah, terutama ketika pola visual antar kelas memiliki kemiripan tinggi dan memerlukan pemahaman konteks global. Hasil 5-*fold cross-validation* menguatkan bahwa keunggulan tersebut stabil di berbagai partisi data yang menunjukkan bahwa kemampuan generalisasi model yang baik meskipun dataset relatif terbatas.

Selain melakukan evaluasi model secara internal, penelitian ini juga membandingkan hasil yang diperoleh dengan temuan dari penelitian-penelitian sebelumnya yang mengkaji klasifikasi citra tanah. Perbandingan ini bertujuan untuk melihat posisi kinerja model yang diusulkan dalam konteks kajian yang lebih luas, sekaligus mengidentifikasi kelebihan dan keterbatasannya dibandingkan metode yang telah ada. Dengan demikian, analisis tidak hanya menekankan performa model pada dataset penelitian ini, tetapi juga memberikan gambaran komprehensif mengenai kontribusi dan relevansi pendekatan yang digunakan dalam perkembangan penelitian klasifikasi citra tanah.

Tabel 13. Perbandingan dengan Penelitian Lain

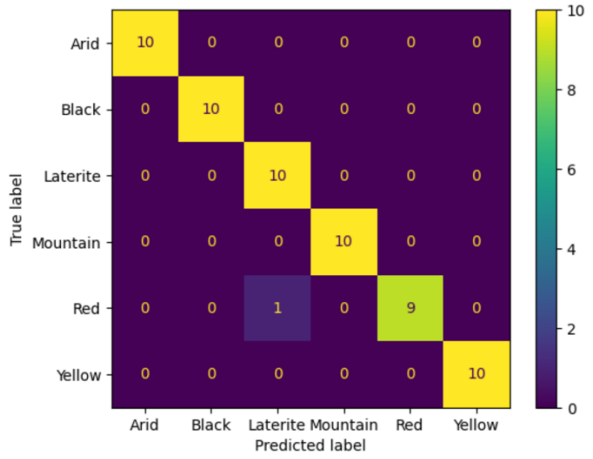
Penulis, Tahun	**Dataset**	**Metode**	**Akurasi (%)**
Barkataki dkk, 2021 [7]	Soil GPR Dataset (7 Kelas)	CNN	97
Lanjewar, 2022 [21]	Soil Dataset (*alluvial, black, clay, and red*)	CNN	97.68
Sheth dkk, 2025 [15]	Soil Dataset (*Alluvial, Arid, Black, Laterite, Mountain, Red,* dan *Yellow*)	CNN+ViTs+adaptive fuzzy logic (original dataset) CNN+ViTs+adaptive fuzzy logic (augmentasi data) CNN+ViTs+adaptive fuzzy logic (CyAUG dataset)	89.32 91.01 100
Darmatasia & Hasrul, 2025	Soil Dataset (*Arid, Black, Laterite, Mountain, Red,* dan *Yellow*)	MobileNet Mobile ViT ViT	66.33 98.67 99.68



Gambar 4. Hasil *Confusion Matrix* Model MobileViT

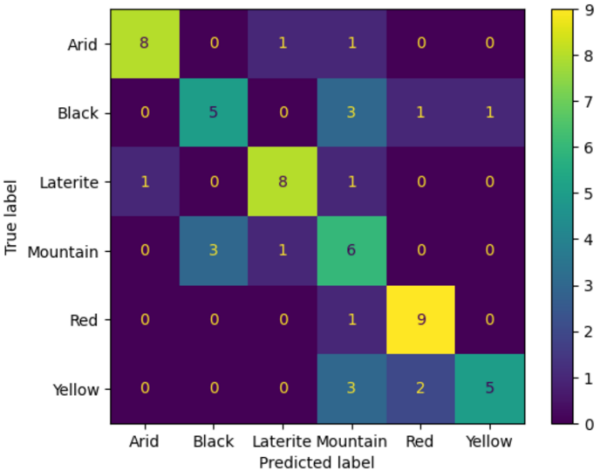
Vol. 10, No. 2, Oktober 2025, Page. 518-533
https://journal.uin-alauddin.ac.id/index.php/instek
p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

Berdasarkan Gambar 4 dapat dilihat bahwa model mampu mengklasifikasikan seluruh sampel dari kelas *Black, Laterite, Mountain, Red, dan Yellow* dengan tingkat akurasi 100%. Namun, untuk kelas *Arid*, model membuat dua kesalahan klasifikasi, di mana dua sampel Arid diprediksi sebagai *Mountain*. Secara keseluruhan, dari total 62 data uji, model berhasil mengklasifikasikan 60 data dengan benar, sehingga menghasilkan tingkat akurasi sekitar 97%. Hal tersebut menunjukkan bahwa model memiliki performa yang sangat baik, meskipun masih terdapat kekeliruan kecil dalam membedakan antara kelas *Arid* dan *Mountain*. Hal tersebut disebabkan karena *Arid* dan *Mountain* menunjukkan karakteristik visual yang cukup mirip sehingga dapat membingungkan model klasifikasi berbasis citra. Tekstur tanah yang retak-retak sangat khas untuk tanah kering atau *Arid Soil* dengan warna dominannya cenderung berada pada spektrum kecokelatan kemerahan yang juga dapat muncul pada jenis tanah seperti *Mountain Soil*. Kedua jenis tanah tersebut sama-sama memiliki rona warna yang berdekatan, didominasi nuansa merah dan cokelat sehingga membuat fitur warna yang merupakan salah satu petunjuk utama bagi model kurang diskriminatif. Kombinasi kemiripan warna, perbedaan skala, serta minimnya fitur tekstur unik membuat kedua jenis tanah tersebut sulit dibedakan, baik oleh model maupun oleh pengamatan visual manusia, terutama dengan kondisi dataset yang tidak memuat variasi kondisi pencahayaan dan tekstur yang cukup untuk membangun representasi fitur yang lebih kuat.



Gambar 5. Hasil *Confusion Matrix* Model *Vision Transformer* Standar

Berdasarkan Gambar 5, dapat dilihat bahwa terdapat 1 salah klasifikasi yang terjadi pada kelas *Red*, di mana 1 citra *Red Soil* diklasifikasikan sebagai *Laterite Soil* yang disebabkan oleh kemiripan visual dengan kelas Laterite. Hal ini dapat menjadi indikasi bahwa beberapa jenis tanah memiliki fitur visual yang tumpang tindih, yang menjadi tantangan umum dalam klasifikasi berbasis citra.



Gambar 6. Hasil *Confusion Matrix* Model MobileNet

Berdasarkan Gambar 6 dapat dilihat bahwa model MobileNet cukup akurat secara keseluruhan, terutama untuk kelas *Red, Arid, dan Laterite*. Namun, masih ada kebingungan antar kelas dengan karakteristik serupa, terutama di kelas *Mountain, Black, dan Yellow*.

4. KESIMPULAN

Penelitian ini mengusulkan penggunaan arsitektur *Mobile Vision Transformer* (MobileViT) sebagai model efisien untuk klasifikasi jenis tanah berbasis citra. Dalam penelitian ini, dilakukan evaluasi terhadap performa tiga arsitektur *deep learning* berbasis transfer *learning*, yaitu *MobileViT*, *Vision Transformer* standar, dan MobileNet. Hasil eksperimen menunjukkan bahwa ViT standar memberikan akurasi klasifikasi tertinggi sebesar 98%, diikuti oleh MobileViT dengan akurasi 97%, dan MobileNet sebesar 68%. Pencapaian akurasi tinggi oleh ViT memerlukan waktu komputasi yang lebih lama, baik pada tahap pelatihan maupun inferensi. Sebaliknya, MobileNet menjadi model dengan waktu komputasi tercepat, meskipun mengorbankan tingkat akurasi yang signifikan. MobileViT memberikan solusi yang seimbang antara akurasi dan efisiensi, dengan akurasi hampir setara ViT namun dengan waktu inferensi yang jauh lebih cepat. MobileViT menunjukkan pengurangan waktu inferensi sekitar 82% dibandingkan Vision Transformer, yaitu dari 783 ms menjadi 141 ms. Hal ini menjadikan MobileViT sangat potensial untuk diterapkan pada perangkat dengan keterbatasan sumber daya komputasi, seperti sistem *edge* atau *mobile*, tanpa menurunkan kualitas klasifikasi secara signifikan. Hasil evaluasi dengan menggunakan *5-fold cross validation* juga menunjukkan bahwa model berbasis Transformer memberikan performa yang tetap konsisten yang menunjukkan bahwa arsitektur tersebut mampu mempertahankan kualitas prediksi pada berbagai partisi data. Temuan dalam penelitian ini menegaskan adanya *trade-off* antara akurasi dan efisiensi komputasi, yang perlu dipertimbangkan dalam pemilihan model berdasarkan konteks penggunaan. Meskipun demikian, penelitian ini secara eksplisit masih dibatasi oleh jumlah dataset yang sangat terbatas sehingga generalisasi model perlu diinterpretasikan dengan hati-hati. Keterbatasan tersebut dapat memengaruhi variasi sampel yang diterima model selama pelatihan. Oleh karena itu, penelitian lanjutan dengan cakupan data yang lebih besar dan lebih beragam diperlukan untuk memvalidasi temuan pada penelitian ini dan memastikan penerapannya pada skenario lapangan yang lebih luas. Penelitian mendatang juga dapat mengeksplorasi pendekatan *fine-tuning* bertahap maupun pengembangan arsitektur *hybrid* yang menggabungkan keunggulan ViT dalam representasi global dan efisiensi *MobileNet* dalam komputasi lokal. Dengan pendekatan yang lebih komprehensif, diharapkan dapat dikembangkan model klasifikasi tanah berbasis citra yang akurat, efisien, dan siap diterapkan secara luas dalam berbagai kebutuhan sistem pertanian dan lingkungan.

DAFTAR PUSTAKA

[1] D. N. Kiran Pandiri, R. Murugan, and T. Goel, “Smart soil image classification system using lightweight convolutional neural network,” *Expert Syst. Appl.*, vol. 238, p. 122185, 2024, doi: https://doi.org/10.1016/j.eswa.2023.122185.

[2] M. Javaid, A. Haleem, I. H. Khan, and R. Suman, “Understanding the potential applications of Artificial Intelligence in Agriculture Sector,” *Adv. Agrochem*, vol. 2, no. 1, pp. 15–30, 2023, doi: https://doi.org/10.1016/j.aac.2022.10.001.

[3] S. Alugoju and P. P, “A Smart Agricultural Framework for Soil Image Classification Using Modified DenseNet and Crop Recommendation System Using Random Forest,” *Int. J. Eng. Trends Technol.*, vol. 72, pp. 119–129, Oct. 2024, doi: 10.14445/22315381/IJETT-V72I10P112.

[4] P. J. and S. T., “Deep learning based multi-labelled soil classification and empirical estimation toward sustainable agriculture,” *Eng. Appl. Artif. Intell.*, vol. 119, p. 105690, 2023, doi: https://doi.org/10.1016/j.engappai.2022.105690.

[5] 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. 2023, doi: 10.3390/su15032374.

[6] M. D. Nguyen *et al.*, “Novel approach for soil classification using machine learning methods,” *Bull. Eng. Geol. Environ.*, vol. 81, no. 11, p. 468, 2022.

[7] N. Barkataki, S. Mazumdar, P. B. D. Singha, J. Kumari, B. Tiru, and U. Sarma, “Classification of soil types from GPR B Scans using deep learning techniques,” in *2021 International*

Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT), 2021, pp. 840–844, doi: 10.1109/RTEICT52294.2021.9573702.

[8] E. K. Gyasi and S. Purushotham, "Advancements in soil classification: an in-depth analysis of current deep learning techniques and emerging trends," *Air, Soil Water Res.*, vol. 16, p. 11786221231214068, 2023.

[9] Y. li, H. Liu, and T. Lv, "A multi-task learning model for global soil moisture prediction based on adaptive weight allocation," *Sci. Rep.*, vol. 15, no. 1, p. 18631, 2025, doi: 10.1038/s41598-025-01894-3.

[10] R. K. Banoth and B. V. R. Murthy, "Soil image classification using transfer learning approach: MobileNetV2 with CNN," *SN Comput. Sci.*, vol. 5, no. 1, p. 199, 2024.

[11] D. Sivabalaselvamani, L. Rahunathan, K. Nanthini, T. Harshini, and C. Hariprasath, "Soil Classification using Deep Learning Techniques," in *2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*, 2023, pp. 582–586, doi: 10.1109/ICAISS58487.2023.10250710.

[12] M. Leonardi and A. Chandra, "Analisis Perbandingan CNN dan Vision Transformer untuk Klasifikasi Biji Kopi Hasil Sangrai," *J. MEDIA Inform. BUDIDARMA*, vol. 8, p. 1398, Jul. 2024, doi: 10.30865/mib.v8i3.7732.

[13] V. Hassija *et al.*, "Transformers for Vision: A Survey on Innovative Methods for Computer Vision," *IEEE Access*, vol. PP, p. 1, Jan. 2025, doi: 10.1109/ACCESS.2025.3571735.

[14] F. Sheth, "Comprehensive Soil Classification Datasets," 2025. .

[15] F. Sheth, P. Mathur, A. K. Gupta, and S. Chaurasia, "An advanced artificial intelligence framework integrating ensembled convolutional neural networks and Vision Transformers for precise soil classification with adaptive fuzzy logic-based crop recommendations," *Eng. Appl. Artif. Intell.*, vol. 158, p. 111425, 2025.

[16] S. Mehta and M. Rastegari, "Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer," *arXiv Prepr. arXiv2110.02178*, 2021.

[17] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv Prepr. arXiv2010.11929*, 2020.

[18] G. Y. Gopal and M. A. Amer, "Mobile vision transformer-based visual object tracking," *arXiv Prepr. arXiv2309.05829*, 2023.

[19] Z. Somogyi, "Performance evaluation of machine learning models," in *The Application of Artificial Intelligence: Step-by-Step Guide from Beginner to Expert*, Springer, 2021, pp. 87–112.

[20] A. Howard *et al.*, "Searching for MobileNetV3," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 1314–1324, doi: 10.1109/ICCV.2019.00140.

[21] M. G. Lanjewar and O. L. Gurav, "Convolutional Neural Networks based classifications of soil images," *Multimed. Tools Appl.*, vol. 81, no. 7, pp. 10313–10336, 2022, doi: 10.1007/s11042-022-12200-y.