# Perbandingan Performa Model EfficientNet-Lite dalam Image Classification

**Disusun Oleh:**

Muhammad Ichsan Firdaus - 140810220025

**PROGRAM STUDI S1 TEKNIK INFORMATIKA**

**FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM**

**UNIVERSITAS PADJADJARAN**

**JATINANGOR**

**2025**

## Abstrak

Penelitian ini membandingkan empat varian model *image classifcation* berbasis EfficientNet-Lite milik TensorFlow: EfficientNet-Lite0 float, EfficientNet-Lite0 int8, EfficientNet-Lite2 float, dan EfficientNet-Lite2 int8. Evaluasi dilakukan terhadap sekumpulan data uji beragam menggunakan metrik *latency* rata-rata, akurasi top-1, akurasi top-3, dan *confidence* rata-rata dari prediksi yang relevan. Hasil menunjukkan adanya trade-off antara kecepatan dan akurasi, serta perbedaan performa antara model float dan *quantized*.

# Pendahuluan

## I. Latar Belakang Masalah

Perkembangan teknologi *artificial intelligence*, khususnya dalam bidang *computer vision*, telah mendorong peningkatan kemampuan sistem dalam mengenali dan memahami gambar digital. Salah satu pendekatan yang banyak digunakan adalah model klasifikasi citra berbasis *deep learning*, yang mampu mengenali objek dalam gambar dengan tingkat akurasi yang tinggi. Model-model ini menjadi inti dari berbagai aplikasi praktis, mulai dari sistem keamanan berbasis wajah, deteksi penyakit dalam dunia medis, hingga pengenalan barang dalam sistem ritel pintar.

TensorFlow sebagai salah satu *framework deep learning* terkemuka menyediakan berbagai model yang telah dilatih dan dioptimalkan untuk berbagai perangkat, termasuk model versi ringan seperti EfficientNet-Lite. Model ini dirancang agar dapat berjalan secara efisien di perangkat dengan sumber daya terbatas seperti *smartphone* atau *edge device* seperti Raspberry pi. TensorFlow menyediakan beberapa varian model EfficientNet-Lite, di antaranya EfficientNet-Lite0 dan EfficientNet-Lite2, yang masing-masing tersedia dalam format kuantisasi float dan int (*integer quantization*).

Meskipun model-model ini telah disiapkan untuk efisiensi dan akurasi, pengguna sering kali menghadapi tantangan dalam memilih model yang paling sesuai dengan kebutuhan aplikasinya. Oleh karena itu, perlu dilakukan studi perbandingan untuk mengevaluasi performa relatif dari masing-masing model berdasarkan berbagai metrik seperti akurasi (top-1 dan top-3), waktu inferensi (*latency*), dan tingkat kepercayaan (*confidence*) terhadap hasil prediksi.

## II. Rumusan Masalah

Berdasarkan latar belakang di atas, rumusan masalah dalam penelitian ini adalah:

1. Model mana yang menunjukkan hasil terbaik dalam hal akurasi top-1 dan top-3?
2. Bagaimana perbandingan waktu inferensi (*latency*) antar model?
3. Apakah terdapat perbedaan signifikan dalam tingkat *confidence* prediksi dari masing-masing model?

4. Bagaimana performa masing-masing model EfficientNet-Lite (Lite0 dan Lite2 dalam versi float dan int) dalam melakukan *image classification* setelah mempertimbangkan ketiga *trade-off* diatas?

## III. Batasan Masalah

Agar penelitian ini lebih fokus dan dapat dilakukan dalam cakupan yang terukur, maka ditetapkan beberapa batasan sebagai berikut:

1. Dataset yang digunakan berupa kumpulan gambar uji (*testing*) dengan label *ground truth* yang telah ditentukan secara manual.
2. Model yang dibandingkan hanya mencakup empat varian: EfficientNet-Lite0 (float), EfficientNet-Lite0 (int), EfficientNet-Lite2 (float), dan EfficientNet-Lite2 (int).
3. Seluruh model merupakan model dasar (*base model*) yang diunduh langsung dari repositori resmi TensorFlow Lite tanpa dilakukan pelatihan ulang (*retraining*), penyesuaian (*fine-tuning*), ataupun optimisasi tambahan lainnya.
4. Klasifikasi dilakukan terhadap citra statis (bukan *real-time video stream*).
5. Penilaian top-k dilakukan dengan pendekatan substring matching, yaitu jika label ground truth merupakan bagian dari prediksi model, maka dianggap benar.
6. Proses inferensi dilakukan dalam lingkungan lokal dengan spesifikasi perangkat keras yang sama untuk semua pengujian.

## IV. Maksud dan Tujuan Penelitian

Penelitian ini bertujuan untuk:

1. Melakukan evaluasi performa empat model *image classification* berbasis EfficientNet-Lite.
2. Membandingkan akurasi top-1 dan top-3 dari masing-masing model terhadap dataset pengujian.
3. Mengukur dan membandingkan rata-rata waktu inferensi (*latency*) pada masing-masing model.
4. Menganalisis tingkat *confidence* prediksi dari model untuk mengetahui seberapa yakin model dalam setiap klasifikasi.

## V. Manfaat Penelitian

Hasil dari penelitian ini diharapkan dapat memberikan manfaat sebagai berikut:

1. Memberikan panduan bagi praktisi atau pengembang sistem dalam memilih model yang tepat untuk perangkat target yang digunakan, dengan mempertimbangkan faktor-faktor utama seperti *model size*, waktu inferensi (*latency*), tingkat akurasi (*accuracy*), serta tingkat kepercayaan prediksi (*confidence*).
2. Menjadi referensi bagi penelitian lanjutan dalam bidang evaluasi model lightweight dan pengaruh teknik kuantisasi terhadap performa inferensi.

## Metode Penelitian

### I. Perangkat dan Lingkungan Eksperimen

Eksperimen dijalankan pada perangkat dengan spesifikasi berikut:

CPU: AMD Ryzen 5 5600H

RAM: 16 GB

Sistem Operasi: Windows

### II. Pengukuran Latency

Latency diukur menggunakan perhitungan waktu mulai sebelum proses inferensi dan waktu selesai setelah model mengembalikan hasil klasifikasi. Nilai latency setiap gambar disimpan dan dirata-ratakan untuk mendapatkan *average latency* tiap model.

Satuan *latency* yang digunakan adalah ms (milisecond)

### III. Pengukuran Akurasi

Setelah model memberikan hasil klasifikasi (berupa daftar top-k kategori teratas), dilakukan evaluasi:

1. Top-1 Accuracy: Benar jika label *ground-truth* merupakan substring dari prediksi urutan pertama.
2. Top-3 Accuracy: Benar jika label ground-truth merupakan substring dari salah satu dari tiga prediksi teratas.

Pendekatan substring matching digunakan karena beberapa label prediksi model terdiri dari dua kata atau lebih (misalnya: mengembalikan TRUE jika ground truth = “dog” dan hasil klasifikasi = "white sheepdog").

### IV. Pengukuran Confidence

Confidence diambil dari skor keyakinan (confidence score) dari setiap prediksi:

1. Confidence dihitung hanya jika label ground-truth termasuk dalam prediksi.
2. Rata-rata confidence dihitung dari confidence score untuk semua prediksi yang benar pada Top-1 dan Top-3.

## V. Evaluasi dan Perbandingan

Setelah semua gambar diuji terhadap seluruh model, dilakukan rekapitulasi hasil per model dalam bentuk:

1. Berat model dalam megabyte (*model size*)
2. Rata-rata waktu inferensi (*average latency*)
3. Persentase top-1 dan top-3 prediksi yang benar (*top-1 acc*, *top-3 acc*)
4. Rata-rata *confidence* dari prediksi yang benar (*average confidence*)

Hasil evaluasi ditampilkan dalam bentuk tabel dan dijadikan dasar pembahasan pada bab selanjutnya.

## Hasil dan Pembahasan

### I. Hasil Perbandingan

| Model | Model Size (MB) | Average Latency (ms) | Top-1 Accuracy (%) | Top-3 Accuracy (%) | Average Confidence |
|---|---|---|---|---|---|
| EfficientNet-Lite0 (int) | 5.3 | 11.82 | 84.00 | 94.00 | 0.62 |
| EfficientNet-Lite0 (float) | 18.1 | 19.12 | 88.00 | 94.00 | 0.66 |
| EfficientNet-Lite2 (int) | 6.9 | 22.35 | 80.00 | 92.00 | 0.49 |
| EfficientNet-Lite2 (float) | 23.7 | 40.76 | 84.00 | 94.00 | 0.52 |

### II. Analisis Hasil

#### 1. Latensi

Model *quantized* (int8) menunjukkan keunggulan signifikan dalam kecepatan inferensi dibanding model float32. EfficientNet-Lite0 (int) mencatat waktu inferensi tercepat, cocok untuk perangkat edge dengan keterbatasan daya dan performa.

#### 2. Akurasi Top-1 dan Top-3

Akurasi model float32 lebih unggul dibanding versi *quantized*. EfficientNet-Lite0 (float) menjadi model paling akurat dalam pengujian, baik untuk prediksi Top-1 maupun Top-3. Hal ini menunjukkan bahwa kompleksitas arsitektur dan presisi data mempengaruhi performa klasifikasi.

#### 3. Confidence

Model float umumnya menghasilkan nilai confidence lebih tinggi, menandakan keyakinan model terhadap prediksinya. Namun, versi quantized juga tidak terlalu jauh tertinggal, tetap memberikan prediksi dengan tingkat kepastian yang cukup baik.

## Kesimpulan

Melihat dari hasil pengujian, didapatkan karakteristik model *quantized* (int8), yaitu memiliki *size* kecil, latensi rendah dengan *trade-off* akurasi dan *confidence* yang tidak kalah jauh dari model float32. Dan untuk model float32 sendiri didapatkan lebih menghasilkan hasil inferensi yang lebih berkualitas.

Meskipun Lite2 (float) memberikan akurasi terbaik, waktu inferensinya dua kali lebih lambat dibanding Lite0 (int). Oleh karena itu, pemilihan model harus mempertimbangkan konteks penggunaan. Untuk aplikasi real-time atau perangkat terbatas, Lite0 (int) bisa lebih sesuai; sementara untuk akurasi tinggi di perangkat dengan sumber daya cukup, Lite2 (float) adalah pilihan ideal.

## Daftar Pustaka

Kaggle. (2017). *ImageNet Object Localization Challenge* [Dataset]. Kaggle. https://www.kaggle.com/c/imagenet-object-localization-challenge/data

Google AI. (2024). *MediaPipe Image Classifier: EfficientNet-Lite models* [Model documentation]. Google AI. https://ai.google.dev/edge/mediapipe/solutions/vision/image_classifier#efficientnet-lite0_model_recommended