

BAB I

PENDAHULUAN

1.1. Latar Belakang

Efisiensi layanan pelanggan dalam industri ritel berperan penting dalam menjaga kepuasan pelanggan (Tan dkk., 2024), terutama di era digital yang semakin kompetitif. Banyak pelanggan memilih belanja *online* untuk menghindari antrian panjang yang memakan waktu (Hamir dkk., 2024). Antrian yang panjang tentu membuat pelanggan frustrasi dan kesal karena waktu berharga mereka terbuang begitu saja hanya untuk mengantri.

Pada skema pembayaran konvensional, umumnya kasir akan memindai *barcode* dari produk satu per satu untuk mengetahui total belanja, namun metode ini memiliki keterbatasan. Proses pemindaian memerlukan interaksi manual, dan sering kali terdapat kendala seperti *barcode* yang sulit ditemukan sehingga menyebabkan waktu transaksi sangat memakan waktu (Kusumawardhani dkk., 2024). Semakin banyak barang yang dibeli, semakin lama waktu pemindaian, sehingga antrian kasir terutama di supermarket besar menjadi tidak terhindarkan.

Beberapa solusi telah dikembangkan untuk mengatasi permasalahan ini, salah satunya adalah Amazon Go yang menggantikan sistem pemindaian *barcode* dengan teknologi *computer vision* dan sensor fusi untuk secara otomatis mendeteksi dan mencatat produk yang dibeli pelanggan, teknologi ini memungkinkan pelanggan untuk langsung keluar

dari toko tanpa perlu melakukan *checkout* manual, yang secara signifikan mengurangi waktu antrian (Hamir dkk., 2024).

Melihat dari keberhasilan Amazon Go, pendekatan berbasis *deep learning* menjadi semakin relevan dalam otomatisasi sistem *checkout*. Mempertimbangkan implementasi pada kehidupan nyata, model deteksi objek yang ideal harus memiliki ukuran yang kecil dan kecepatan inferensi yang tinggi. Model dengan kriteria seperti itu dapat mengurangi konsumsi daya selama proses inferensi yang pada akhirnya dapat berdampak pada efisiensi biaya operasional, terutama dalam sistem *checkout* otomatis yang harus berjalan secara *real-time* dalam jangka waktu yang panjang.

Pada konteks ini, model YOLO (*You Only Look Once*) menjadi salah satu pilihan populer karena dikenal memiliki keseimbangan antara kecepatan dan akurasi dalam deteksi objek. Jika dibandingkan dengan metode deep learning lainnya yang menggunakan *two stage detectors* seperti R-CNN, Fast R-CNN, Faster R-CNN, FPN, Mask R-CNN, dan SPP, model YOLO yang menggunakan *one stage detector* lebih cepat pemrosesannya dikarenakan memproses gambar dalam sekali jalan (More & Bhosale, 2023).

Pada versi terbarunya, menurut penelitian Rasheed & Zarkoosh (2024), meskipun YOLOv11 telah menunjukkan efisiensi dalam aspek akurasi, ukuran model, waktu inferensi, dan konsumsi daya baik dengan *pruning* atau tidak, model ini masih menggunakan Float 16 (FP16) untuk mempresentasikan parameter model, memiliki kompleksitas komputasi

lebih tinggi sehingga waktu inferensi lebih lama. Oleh karena itu, mereka merekomendasikan untuk melakukan eksperimen dengan kuantisasi ke *bit* yang lebih kecil di penelitian selanjutnya untuk mengurangi ukuran model dan meringankan beban komputasi sehingga mempercepat waktu inferensi dan penggunaan memori akan lebih efisien (Hawks dkk., 2021). Optimasi seperti ini penting untuk membuka peluang model YOLOv11 digunakan secara lebih luas, terutama pada *edge devices* atau perangkat keras dengan spesifikasi terbatas, sehingga meningkatkan fleksibilitas implementasi di berbagai lingkungan pengembangan.

Berbagai penelitian terkait kuantisasi telah dilakukan sebelumnya, seperti pada penelitian (Liu dkk., 2024) model YOLOv5 mereka dikuantisasi pada bagian tertentu di arsitekturnya yang difokuskan untuk diimplementasikan di *edge devices*. Hasilnya terbukti bahwa kuantisasi dapat menurunkan kompleksitas komputasi dan memperkecil ukuran model, namun kuantisasi membuat akurasi model menurun secara signifikan. Sedangkan pada penelitian Bouzekraoui dkk., (2024) yang difokuskan untuk mendeteksi objek secara real time pada marka jalan dengan YOLOv8 yang sudah dikuantisasi ke bentuk format INT8, hasilnya terbukti dengan peningkatan FPS, yang artinya kuantisasi dapat meningkatkan efisiensi dan kinerja yang signifikan saat dilakukan inferensi pada perangkat yang menggunakan CPU kelas bawah.

Sementara itu, penelitian terkait model YOLO untuk retail juga telah dilakukan, seperti pada penelitian Hamir dkk., (2024) yang menggunakan

YOLOv5 untuk deteksi objek ritel dalam kasus checkout otomatis, dengan hasil model mereka tidak mendeteksi cukup baik saat diberikan gambar yang baru dan random, karena model overfitting sehingga tidak memahami pola pada objeknya. Pada penelitian Wibowo, (2024) dihasilkan model dengan mAP50 sebesar 0.98 (98%) dan mampu mendeteksi produk ritel dengan baik, namun waktu inferensinya masih cukup lama yaitu sekitar 4.9 detik tercepatnya sehingga dibutuhkan optimasi lebih lanjut. Pada penelitian Andaru, (2024), yang menggunakan YOLOv8s untuk mendeteksi objek ritel di rak. Hasil penelitian tersebut menunjukkan akurasi sebesar 94.7% dengan waktu inferensi yang lebih cepat saat menggunakan YOLOv8n (versi nano). Pada penelitian (Chhabra & Goyal, 2024) mereka menggunakan YOLO-NAS yang ditingkatkan menggunakan hyperparameter tuning dalam kasus deteksi objek ritel untuk tunanetra mendapatkan hasil mAP50 sebesar 96.80% untuk dataset Grozi-120 dan 97.61% untuk dataset Retail Product yang diambil dari roboflow. Selain itu, penggunaan kuantisasi pada model HYOLO-NAS mereka dapat mempercepat proses eksekusi model sehingga cocok untuk dijalankan pada perangkat dengan sumber daya terbatas.

Berdasarkan kajian dari penelitian terkait, menunjukkan bahwa kuantisasi dapat memperkecil model serta mempercepat waktu inferensinya, namun belum banyak diterapkan pada YOLOv11 untuk kasus deteksi produk ritel FMCG (*Fast-Moving Consumer Goods*) dengan kategori makanan dan minuman. Kategori produk tersebut dipilih karena

visual yang beragam, ketersediaan data, dan referensi yang melimpah dari studi sebelumnya.

1.2.Rumusan Masalah

Berdasarkan latar belakang yang sudah dipaparkan sebelumnya, rumusan masalah penelitian ini adalah sebagai berikut :

1. Bagaimana pemilihan konfigurasi hyperparameter yang optimal dapat meningkatkan performa model YOLOv11 dalam mendeteksi objek ritel FMCG?
2. Bagaimana pengaruh kuantisasi model YOLOv11 terhadap performa model dalam mendeteksi objek ritel FMCG?

1.3.Tujuan Penelitian

Berdasarkan rumusan masalah yang telah diuraikan, penelitian ini bertujuan untuk :

1. Menganalisis dan mengevaluasi dampak kuantisasi terhadap model YOLOv11 dalam mendeteksi objek ritel berdasarkan akurasi dan kecepatan inferensi.
2. Mengimplementasikan konfigurasi hyperparameter yang sesuai untuk meningkatkan performa deteksi objek ritel dengan YOLOv11.

1.4.Manfaat Penelitian

1. Efisiensi Komputasi

Model yang dihasilkan ukurannya lebih kecil, dan waktu inferensi lebih cepat karena kompleksitas komputasinya berkurang. Sehingga

perangkat yang menggunakan CPU kelas rendah bahkan tanpa bantuan GPU bisa menjalankannya dengan cepat.

2. Kontribusi Pada Penelitian AI Lokal

Dataset custom yang dikembangkan pada penelitian ini dapat menjadi kontribusi bagi pengembangan AI di Indonesia, khususnya dalam deteksi objek produk ritel. Dataset ini juga dapat digunakan untuk penelitian-penelitian serupa di masa depan.

3. Dukungan Terhadap Perangkat Terbatas

Ukuran model yang lebih kecil memungkinkan implementasi pada perangkat dengan spesifikasi terbatas. Hal ini membuka peluang untuk memperluas penerapan model deteksi objek ritel di berbagai lingkungan, sekaligus menjadi dasar penelitian lanjutan untuk meningkatkan efisiensi di perangkat dengan sumber daya minimal.

1.5.Batasan Masalah

1. Berfokus pada pengembangan model deteksi objek produk ritel FMCG (*Fast-Moving Consumer Goods*) kategori makanan dan minuman dengan YOLOv11.
2. Pengambilan dataset dilakukan secara manual untuk memberikan fleksibilitas dalam mengontrol kualitas data. Metode ini memungkinkan untuk melakukan penyesuaian variasi sudut pengambilan gambar, latar belakang, serta kondisi pencahayaan. Jika ditemukan kekurangan dalam variasi data, dataset dapat diperbarui dengan mudah atau ditambahkan sesuai kebutuhan guna meningkatkan kualitas hasil deteksi.

3. Dataset custom yang digunakan dibatasi menjadi hanya 10 produk gambar produk ritel FMCG makanan dan minuman yang ada di Indonesia dengan total 1.000 gambar agar tidak terlalu membebani perangkat yang digunakan, diambil secara manual menggunakan perangkat android dan dianotasi manual menggunakan Roboflow serta diset untuk publik agar dapat diakses dan diverifikasi oleh pihak lain jika dibutuhkan.
4. Penelitian ini tidak memperhitungkan variasi ukuran benda, data yang digunakan diasumsikan bahwa setiap objek mewakili kategori yang spesifik tanpa memandang ukuran fisik objek tersebut
5. Fokus penelitian adalah pada kecepatan inferensi, ukuran model, dan performa deteksi tanpa membahas aspek keamanan sistem dan antarmuka secara detail.
6. Kuantisasi model dilakukan dengan konversi *bit* model ke Integer 8 (INT8) menggunakan ONNX Runtime tanpa mencakup teknik kuantisasi lain.