KOMPARASI DYNAMIC RANGE DAN FLOAT16 QUANTIZATION PADA MODEL MOBILEFACENET UNTUK VERIFIKASI WAJAH MENGGUNAKAN TENSORFLOW LITE

Ferry Hasan[1], I Gede Susrama Mas Diyasa[2*], Achmad Junaidi[3]
[1,2,3] Program Studi Informatika, Universitas Pembangunan Nasional "Veteran" Jawa Timur, Surabaya, Indonesia

**Penulis korespondensi: igsusrama.if@upnjatim.ac.id*

ABSTRAK

Pengenalan wajah berbasis kecerdasan buatan telah banyak diadopsi dalam aplikasi presensi dan autentikasi digital, namun implementasinya pada perangkat *mobile* kelas menengah ke bawah masih menghadapi tantangan efisiensi yang signifikan. Penelitian ini menerapkan pendekatan *post-training quantization*, yaitu teknik kompresi model yang dilakukan setelah proses pelatihan selesai tanpa memerlukan pelatihan ulang, untuk mengoptimalkan model MobileFaceNet agar dapat dijalankan secara efisien pada perangkat dengan sumber daya komputasi terbatas. Dua metode kuantisasi dari kerangka kerja TensorFlow Lite dievaluasi secara komparatif, yaitu *dynamic range quantization* yang mengonversi bobot model dari *32-bit floating-point* menjadi *8-bit integer*, serta *float16 quantization* yang mengonversi bobot menjadi *16-bit floating-point*. Model dilatih menggunakan dataset CASIA-WebFace dengan fungsi kerugian ArcFace dan dievaluasi menggunakan protokol verifikasi standar *Labeled Faces in the Wild* (LFW) dengan skema *10-fold cross validation* terhadap 6.000 pasang citra wajah. Hasil eksperimen menunjukkan bahwa *dynamic range quantization* menghasilkan rasio kompresi sebesar 3,32× dari 3,83 MB menjadi 1,15 MB, sedangkan *float16 quantization* menghasilkan rasio kompresi sebesar 1,96× menjadi 1,95 MB. Model *baseline* mencapai akurasi 90,88%, *dynamic range* 90,83%, dan *float16* 90,90% dengan penurunan akurasi relatif <1% terhadap *baseline*. Ketiga model memiliki nilai *Area Under Curve* (AUC) sebesar 0,958 yang menunjukkan kemampuan diskriminasi setara. Temuan ini mengindikasikan bahwa *post-training quantization*, khususnya *dynamic range* untuk kompresi maksimal dan *float16* untuk stabilitas akurasi, merupakan strategi kompresi yang efektif untuk implementasi pengenalan wajah pada perangkat *mobile* kelas menengah ke bawah.

Kata kunci: post-training quantization, MobileFaceNet, dynamic range quantization, float16 quantization, TensorFlow Lite.

1 PENDAHULUAN

Pengenalan wajah berbasis kecerdasan buatan telah banyak diadopsi dalam aplikasi presensi digital maupun autentikasi pengguna pada lingkungan pendidikan dan industri. Mayoritas pengguna *smartphone* di Indonesia masih menggunakan perangkat pada segmen harga terjangkau, sehingga pengembangan sistem pengenalan wajah yang dapat berjalan efisien pada perangkat kelas menengah ke bawah memiliki relevansi praktis yang tinggi, terutama pada perangkat dengan chipset seperti Qualcomm Snapdragon 6 *series* yang umum digunakan pada segmen tersebut (*Qualcomm Hexagon NPU*, t.t.). Perangkat pada kelas ini memiliki keterbatasan kapasitas memori dan unit akselerasi perangkat keras dibandingkan chipset kelas atas, sehingga model dengan presisi float32 penuh memakan alokasi RAM yang besar dan kurang efisien untuk diimplementasikan

secara langsung. Oleh karena itu, optimasi ukuran model dan efisiensi kapasitas penyimpanan (*memory footprint*) menjadi langkah fundamental sebelum implementasi sistem presensi dilakukan.

Model pengenalan wajah berbasis *deep learning* umumnya memiliki kompleksitas komputasi yang tinggi sehingga sulit diimplementasikan secara langsung pada perangkat dengan sumber daya terbatas (Wang & Deng, 2021). MobileFaceNet diusulkan sebagai arsitektur ringan dengan parameter di bawah satu juta yang menggantikan *Global Average Pooling* dengan *Global Depthwise Convolution* untuk mempertahankan informasi spasial penting bagi diskriminasi wajah (Chen dkk., 2018). Untuk meningkatkan kemampuan diskriminatif pada ruang *embedding*, ArcFace Loss menambahkan margin sudut aditif yang memaksa fitur antar kelas saling terpisah lebih jelas dibandingkan fungsi *Softmax* konvensional (Deng dkk., 2022). Beberapa penelitian lanjutan mencoba meningkatkan efisiensi arsitektur MobileFaceNet, seperti modifikasi struktur layer dan modul atensi yang dilakukan oleh Xiao dkk. (2022), serta penggunaan teknik optimasi pelatihan *Sharpness-Aware Minimization* oleh Hassanpour & Kowsari (2023). Pendekatan arsitektur ringan lainnya turut dikembangkan melalui kombinasi CNN dan transformer ringan pada EdgeFace (George dkk., 2024), pencarian arsitektur otomatis pada PocketNet (F. Boutros dkk., 2022), serta deteksi dan penyelarasan wajah multi-tahap pada MTCNN yang banyak digunakan sebagai komponen *preprocessing* standar (K. Zhang dkk., 2016).

Meskipun optimasi arsitektur menunjukkan hasil yang menjanjikan, sebagian besar penelitian tersebut masih diuji pada lingkungan komputasi berbasis GPU desktop dan belum mengevaluasi implementasi langsung pada perangkat *mobile* dengan spesifikasi terbatas. Di sisi lain, pendekatan kuantisasi sebagai alternatif kompresi model juga telah mendapatkan perhatian dalam literatur. Boutros dkk. (2022) mengajukan QuantFace yang memanfaatkan kuantisasi bit rendah berbasis data sintetis, namun pendekatan tersebut tetap memerlukan proses pelatihan ulang (*quantization-aware training*) dan penggunaan dataset tambahan sebagai representasi kalibrasi, sehingga tidak sepenuhnya bersifat *plug and play*. B. Jacob dkk. (2018) meletakkan fondasi teoritis bagi inferensi integer-only pada jaringan saraf, namun fokus penelitiannya adalah pada kuantisasi berbasis pelatihan, bukan kuantisasi pasca-pelatihan yang langsung dapat diterapkan setelah model selesai dilatih. Sebagai alternatif optimasi yang tidak memerlukan perubahan arsitektur maupun pelatihan ulang, pendekatan *post-training quantization* menawarkan solusi yang lebih sederhana dan mudah direproduksi tanpa memerlukan dataset kalibrasi tambahan (Boutros dkk., 2022). TensorFlow Lite menyediakan beberapa metode *post-training quantization*, di antaranya *dynamic range quantization* yang mengonversi bobot ke representasi *8-bit integer* dan *float16 quantization* yang mengonversi bobot ke representasi *16-bit floating-point*, dengan karakteristik berbeda terhadap ukuran model dan latensi inferensi (*Post-Training Quantization* | Google AI Edge, 2026). Namun, hingga saat ini belum ditemukan penelitian yang secara sistematis membandingkan efektivitas kedua metode tersebut secara head-to-head pada model MobileFaceNet dengan menggunakan protokol evaluasi standar LFW, sehingga terdapat celah penelitian yang perlu diisi.

Performa model pada perangkat mobile juga sangat dipengaruhi oleh spesifikasi perangkat keras yang digunakan, sehingga pengujian langsung pada perangkat target menjadi penting untuk memperoleh hasil yang representatif (Li dkk., 2020). Berdasarkan uraian tersebut, penelitian ini bertujuan menerapkan dan membandingkan dynamic range quantization dan float16 quantization pada model MobileFaceNet menggunakan kerangka kerja TensorFlow Lite, serta mengevaluasi pengaruhnya terhadap akurasi verifikasi wajah dan ukuran model berdasarkan protokol LFW dengan skema 10-fold cross validation. Penelitian ini membatasi ruang lingkup pada evaluasi kompresi arsitektural secara offline sebelum tahap *deployment* ke perangkat mobile. Pembatasan

ini dilakukan secara sengaja karena tujuan utama penelitian adalah mengukur degradasi akurasi biometrik akibat penurunan presisi numerik bobot model, yang merupakan properti intrinsik model itu sendiri dan tidak bergantung pada perangkat keras tertentu. Evaluasi latensi dan efisiensi sumber daya pada perangkat fisik merupakan tahap validasi lanjutan yang disarankan sebagai penelitian berikutnya, setelah kelayakan akurasi model terkompresi terbukti terlebih dahulu melalui protokol evaluasi standar.

2 METODE

2.1 Dataset Dan Preprocessing

Gambar 1. Diagram alur metodologi penelitian

Penelitian ini menggunakan dua dataset publik. Dataset CASIA-WebFace digunakan sebagai data pelatihan dengan 489.839 citra dari 10.572 identitas yang berhasil lolos deteksi wajah, sedangkan dataset *Labeled Faces in the Wild* (LFW) digunakan sebagai data evaluasi dengan 13.233 citra dari 5.749 identitas yang seluruhnya berhasil lolos deteksi (Wang & Deng, 2021). Setiap citra diproses melalui tahap deteksi dan penyelarasan wajah menggunakan *Multi-task Cascaded Convolutional Networks* (MTCNN) untuk memperoleh *bounding box* dan titik *landmark* wajah (K. Zhang dkk., 2016). Citra hasil deteksi kemudian diubah ukurannya menjadi 112×112 piksel sesuai kebutuhan input MobileFaceNet, dan nilai pikselnya dinormalisasi ke rentang -1 hingga 1 agar distribusi input lebih stabil selama pelatihan.

2.2 Pelatihan Model Baseline

Model MobileFaceNet dilatih menggunakan arsitektur sesuai spesifikasi Chen dkk. (2018) yang menghasilkan vektor *embedding* berdimensi 128 setelah dinormalisasi secara L2. Arsitektur ini mengadaptasi struktur *inverted residual* milik MobileNetV2 dengan faktor ekspansi yang lebih kecil pada lapisan *bottleneck* untuk menjaga jumlah parameter tetap rendah (Sandler dkk., 2019). Fungsi kerugian yang digunakan adalah ArcFace Loss dengan margin sudut 0,5 dan skala 64 sesuai konfigurasi yang direkomendasikan (Deng dkk., 2022). Proses pelatihan menggunakan optimizer *Stochastic Gradient Descent* dengan jadwal *learning rate warmup cosine decay* dan mekanisme *gradient clipping* untuk menjaga stabilitas konvergensi. Mekanisme *early stopping* berbasis nilai *validation loss* diterapkan dengan toleransi 15 *epoch* untuk mencegah *overfitting.*

2.3 Penerapan Post-Training Quantization

Sebelum proses kuantisasi dilakukan, *layer* ArcFace dipisahkan dari model karena hanya diperlukan untuk perhitungan *loss* selama pelatihan, sehingga yang dikonversi adalah *backbone* MobileFaceNet yang menghasilkan *embedding* 128 dimensi. Konversi dilakukan menggunakan TFLite Converter untuk menghasilkan tiga varian model, yaitu model *baseline* dalam format *float32* tanpa kuantisasi, model hasil *dynamic range quantization* yang mengonversi bobot menjadi representasi *8-bit integer* dengan aktivasi tetap dihitung dalam *float32* secara dinamis,

dan model hasil *float16 quantization* yang mengonversi bobot menjadi representasi 16-*bit floating-point* (B. Jacob dkk., 2018). Kedua metode kuantisasi tidak memerlukan *representative dataset* untuk kalibrasi, sehingga proses konversi dapat dilakukan langsung setelah pelatihan selesai tanpa pelatihan ulang (Boutros dkk., 2022). Setiap model hasil konversi diverifikasi menggunakan *dummy input* untuk memastikan bentuk *output* 1×128 dan nilai L2 *norm* mendekati 1,0, yang menunjukkan bahwa lapisan normalisasi tetap berfungsi setelah proses kuantisasi.

2.4 Evaluasi LFW 10-Fold Cross Validation

Evaluasi akurasi dilakukan menggunakan protokol standar LFW dengan 6.000 pasang citra wajah yang dibagi menjadi 10-*fold*, terdiri atas 3.000 pasangan *same-person* dan 3.000 pasangan *different-person*. Pada setiap iterasi, sembilan *fold* digunakan untuk menentukan *threshold* optimal berdasarkan nilai *cosine similarity*, sedangkan satu *fold* sisanya digunakan sebagai data uji secara bergantian. Metrik yang dihitung meliputi akurasi verifikasi, *False Acceptance Rate* (FAR), *False Rejection Rate* (FRR), *Equal Error Rate* (EER), dan *Area Under Curve* (AUC) dari kurva *Receiver Operating Characteristic*. Penurunan akurasi relatif terhadap *baseline* dihitung untuk menentukan apakah model hasil kuantisasi masih dapat diterima secara praktis, dengan ambang toleransi di bawah 1% mengacu pada praktik umum penelitian kompresi model biometrik (Li dkk., 2020).

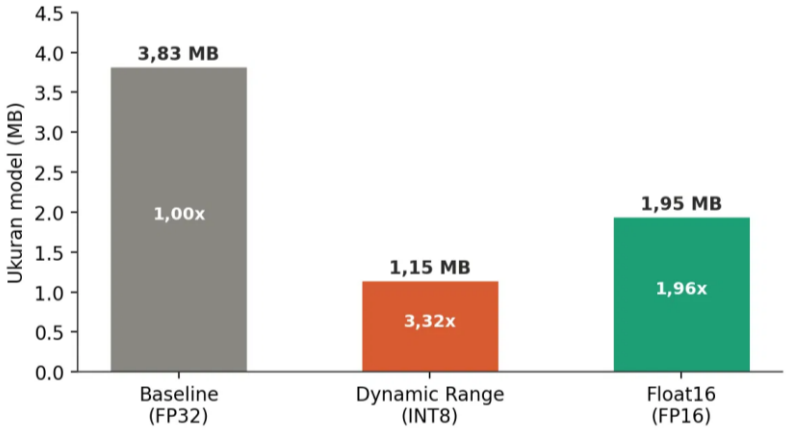
3 HASIL DAN PEMBAHASAN

3.1 Hasil Preprocessing dan Pelatihan Model

Dari 494.414 citra mentah pada dataset CASIA-WebFace, sebanyak 489.839 citra berhasil lolos deteksi dan penyelarasan wajah menggunakan MTCNN, sedangkan seluruh 13.233 citra pada dataset LFW berhasil terdeteksi tanpa kegagalan, yang mengindikasikan kualitas dataset evaluasi yang konsisten (K. Zhang dkk., 2016). Proses pelatihan berlangsung selama 72 *epoch* sebelum dihentikan oleh mekanisme *early stopping*, dengan model terbaik diperoleh pada *epoch* ke-57 berdasarkan nilai *validation loss* terendah sebesar 19,7764. Pola penurunan *training loss* dan *validation loss* yang konsisten menunjukkan bahwa proses optimasi berlangsung stabil tanpa indikasi *overfitting* yang signifikan.

3.2 Hasil Penerapan Post-Training Quantization

Ketiga varian model berhasil dikonversi ke format TensorFlow Lite dengan hasil verifikasi *output* yang konsisten, yaitu bentuk *embedding* 1×128 dengan nilai L2 *norm* sebesar 1,000000 pada seluruh model. Perbandingan ukuran file dan rasio kompresi ketiga model ditunjukkan pada Gambar 2 berikut.



Gambar 2. Perbandingan ukuran file dan rasio kompresi model

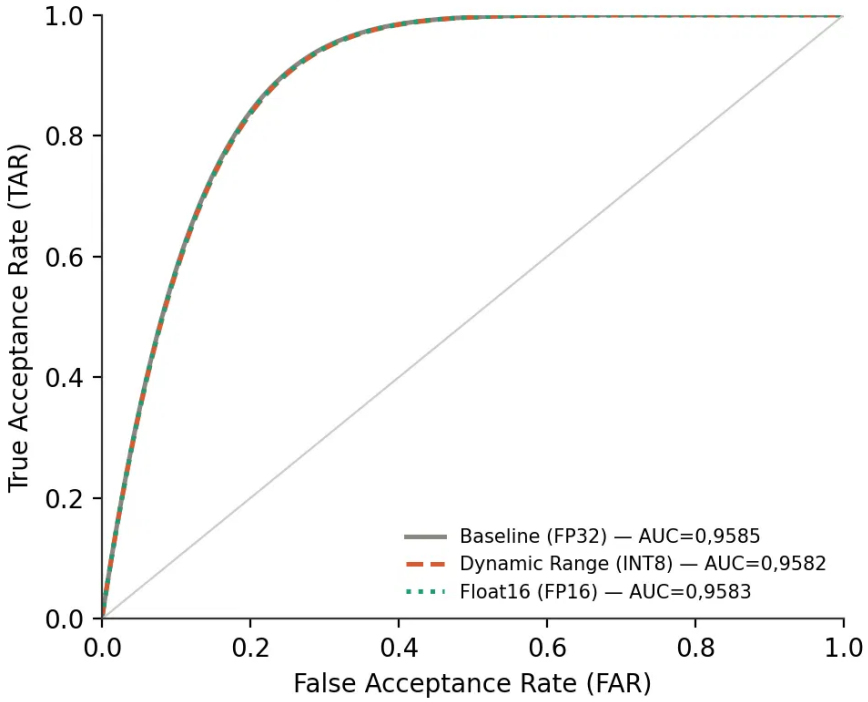
Berdasarkan Gambar 2, *dynamic range quantization* menghasilkan rasio kompresi tertinggi sebesar 3,32 kali, mendekati nilai teoritis 4 kali yang diharapkan dari konversi representasi 4 *byte* menjadi 1 *byte* per parameter. Selisih dari nilai teoritis disebabkan oleh metadata dan struktur *graph* yang tidak seluruhnya mengalami kuantisasi (*Post-Training Quantization | Google AI Edge*, 2026). Sementara itu, *float16 quantization* menghasilkan rasio kompresi 1,96 kali, sangat mendekati nilai teoritis 2 kali. Temuan ini sejalan dengan karakteristik kedua metode yang dijelaskan dalam dokumentasi kerangka kerja TensorFlow Lite, menunjukkan representasi bit yang lebih rendah secara konsisten menghasilkan kompresi yang lebih agresif (Li dkk., 2020).

3.3 Hasil Evaluasi Akurasi LFW 10-Fold Cross Validation

Hasil evaluasi ketiga model pada dataset LFW menggunakan protokol 10-*fold cross validation* ditunjukkan pada Tabel 1 berikut.

Tabel 1. Hasil evaluasi LFW 10-fold cross validation

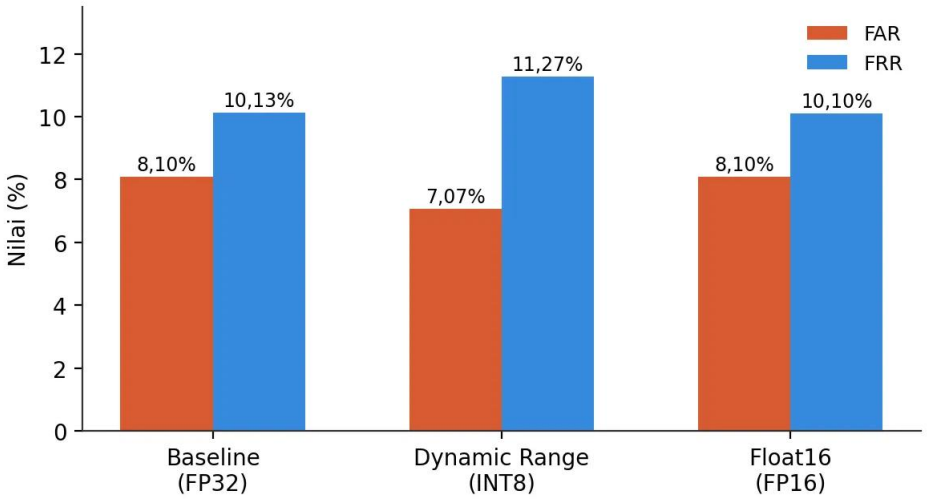
Model	Akurasi	FAR	FRR	EER	AUC
Baseline (Float32)	90,88% ± 1,12	8,10%	10,13%	9,27%	0,9585
Dynamic Range (INT8)	90,83% ± 1,17	7,07%	11,27%	9,33%	0,9582
Float16 (FP16)	90,90% ± 1,12	8,10%	10,10%	9,27%	0,9583



Gambar 3. Kurva ROC ketiga varian model pada dataset LFW

Berdasarkan Tabel 1, ketiga model menunjukkan performa yang sangat berdekatan, dengan selisih akurasi terhadap *baseline* berada di bawah 1%, yaitu 0,05 *percentage point* untuk *dynamic range* dan 0,02 *percentage point* untuk *float16*. Hasil ini menunjukkan bahwa penurunan presisi numerik pada representasi bobot tidak menyebabkan degradasi akurasi yang berarti, sejalan dengan temuan pada penelitian kompresi model wajah lainnya yang melaporkan stabilitas akurasi

serupa setelah penerapan teknik kuantisasi (Boutros dkk., 2022). Nilai *Area Under Curve* yang hampir identik pada ketiga model, yaitu berkisar 0,958, mengindikasikan bahwa kemampuan diskriminasi antara pasangan wajah *same-person* dan *different-person* tidak terpengaruh secara fundamental oleh proses kuantisasi. Meskipun akurasi keseluruhan relatif setara, terdapat perbedaan karakteristik pada nilai FAR dan FRR antar model. Model *dynamic range* menghasilkan FAR yang lebih rendah dibandingkan *baseline*, namun FRR yang lebih tinggi, yang mengindikasikan model ini cenderung lebih konservatif dalam menerima pasangan wajah sebagai identitas yang sama. Sebaliknya, model *float16* menghasilkan nilai FAR dan FRR yang hampir identik dengan *baseline*, menunjukkan bahwa representasi 16-*bit floating-point* mampu mempertahankan karakteristik keputusan model asli secara lebih dekat dibandingkan representasi *integer 8-bit*. Pola ini konsisten dengan karakteristik teoritis kedua metode, *float16* mempertahankan presisi numerik yang lebih tinggi karena tetap berada dalam representasi *floating-point* (Li dkk., 2020).



Gambar 4. Perbandingan FAR dan FRR rata-rata 10-fold

Gambar 4 secara visual mempertegas perbedaan karakteristik antar model pada dimensi FAR dan FRR. Terlihat bahwa model *dynamic range* (INT8) memiliki batang FAR yang paling rendah (7,07%) namun batang FRR yang paling tinggi (11,27%) di antara ketiga model. Kondisi ini menunjukkan bahwa kuantisasi ke representasi *integer 8-bit* menggeser titik keputusan model ke arah yang lebih ketat, sehingga lebih sedikit pasangan wajah yang berbeda diterima sebagai sama (FAR rendah), tetapi lebih banyak pasangan wajah yang sebenarnya sama ditolak (FRR tinggi). Sebaliknya, model *float16* (FP16) menunjukkan profil FAR dan FRR yang hampir identik dengan model *baseline* (FP32), yaitu masing-masing 8,10% dan 10,10%, yang mengonfirmasi bahwa presisi *16-bit floating-point* cukup untuk menjaga perilaku keputusan model tetap mendekati versi aslinya.

Berdasarkan kombinasi hasil rasio kompresi dari Gambar 2 dan hasil akurasi dari Tabel 1, penelitian ini mengidentifikasi adanya *trade-off* yang jelas antara kedua metode kuantisasi. *Dynamic range quantization* lebih sesuai digunakan pada skenario yang memprioritaskan efisiensi ukuran model dan penghematan memori secara maksimal, sedangkan *float16 quantization* lebih sesuai digunakan pada skenario yang memprioritaskan stabilitas karakteristik keputusan model

mendekati *baseline*. Temuan ini melengkapi penelitian Xiao dkk., (2022) dan Hassanpour & Kowsari, (2023) yang berfokus pada optimasi arsitektur, dengan menunjukkan bahwa optimasi pasca-pelatihan tanpa perubahan struktur model maupun pelatihan ulang tetap dapat menghasilkan kompresi yang signifikan tanpa mengorbankan kemampuan verifikasi wajah secara berarti. Hasil ini relevan untuk implementasi praktis pada aplikasi presensi atau autentikasi berbasis wajah di perangkat *mobile* dengan keterbatasan memori dan daya komputasi (Duong dkk., 2019).

4 KESIMPULAN

Penelitian ini menunjukkan bahwa penerapan *post-training quantization* pada model MobileFaceNet menggunakan TensorFlow Lite berhasil mengurangi ukuran model secara signifikan tanpa mengorbankan akurasi verifikasi wajah secara berarti. *Dynamic range quantization* menghasilkan rasio kompresi tertinggi sebesar 3,32 kali dengan akurasi 90,83%, sedangkan *float16 quantization* menghasilkan rasio kompresi 1,96 kali dengan akurasi 90,90% yang paling mendekati *baseline* sebesar 90,88%. Kedua metode mempertahankan penurunan akurasi relatif di bawah 1% dan nilai *Area Under Curve* yang setara pada kisaran 0,958, sehingga keduanya layak digunakan sebagai strategi optimasi untuk implementasi pengenalan wajah pada perangkat *mobile* kelas menengah ke bawah. *Dynamic range quantization* lebih sesuai dipilih ketika efisiensi ukuran model menjadi prioritas utama, sedangkan *float16 quantization* lebih sesuai dipilih ketika stabilitas karakteristik keputusan model terhadap *baseline* menjadi pertimbangan utama. Penelitian lanjutan disarankan untuk mengevaluasi performa kedua model hasil kuantisasi secara langsung pada perangkat Android dalam berbagai kondisi pencahayaan, guna memperoleh gambaran komprehensif mengenai *trade-off* antara akurasi, latensi inferensi, dan efisiensi sumber daya pada skenario penggunaan nyata

UCAPAN TERIMAKASIH

Penulis mengucapkan terima kasih kepada Program Studi Informatika, Fakultas Ilmu Komputer, Universitas Pembangunan Nasional “Veteran” Jawa Timur atas dukungan fasilitas dan bimbingan akademik oleh dosen pembimbing selama penelitian ini berlangsung. Selain itu, penulis menyampaikan apresiasi kepada pihak-pihak yang telah memberikan izin untuk membantu proses pengembangan pada penelitian ini.

DAFTAR PUSTAKA

B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, & D. Kalenichenko. (2018). Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2704–2713. https://doi.org/10.1109/CVPR.2018.00286

Boutros, F., Damer, N., & Kuijper, A. (2022). *QuantFace: Towards Lightweight Face Recognition by Synthetic Data Low-bit Quantization* (arXiv:2206.10526). arXiv. https://doi.org/10.48550/arXiv.2206.10526

Chen, S., Liu, Y., Gao, X., & Han, Z. (2018). MobileFaceNets: Efficient CNNs for Accurate Real-Time Face Verification on Mobile Devices. Dalam J. Zhou, Y. Wang, Z. Sun, Z. Jia, J. Feng, S. Shan, K. Ubul, & Z. Guo (Ed.), *Biometric Recognition* (hlm. 428–438). Springer International Publishing.

Deng, J., Guo, J., Yang, J., Xue, N., Kotsia, I., & Zafeiriou, S. (2022). ArcFace: Additive Angular Margin Loss for Deep Face Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*, 5962–5979. https://doi.org/10.1109/TPAMI.2021.3087709

Duong, C. N., Quach, K. G., Jalata, I., Le, N., & Luu, K. (2019). MobiFace: A Lightweight Deep Learning Face Recognition on Mobile Devices. *2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 1–6. https://doi.org/10.1109/BTAS46853.2019.9185981

F. Boutros, P. Siebke, M. Klemt, N. Damer, F. Kirchbuchner, & A. Kuijper. (2022). PocketNet: Extreme Lightweight Face Recognition Network Using Neural Architecture Search and Multistep Knowledge Distillation. *IEEE Access*, *10*, 46823–46833. https://doi.org/10.1109/ACCESS.2022.3170561

George, A., Ecabert, C., Shahreza, H. O., Kotwal, K., & Marcel, S. (2024). *EdgeFace: Efficient Face Recognition Model for Edge Devices* (arXiv:2307.01838). arXiv. https://doi.org/10.48550/arXiv.2307.01838

Hassanpour, A., & Kowsari, Y. (2023). Lightweight Face Recognition: An Improved MobileFaceNet Model. *arXiv:2311.15326 [cs.CV]*. http://arxiv.org/abs/2311.15326

K. Zhang, Z. Zhang, Z. Li, & Y. Qiao. (2016). Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Processing Letters*, *23*(10), 1499–1503. https://doi.org/10.1109/LSP.2016.2603342

Li, H. C., Deng, Z. Y., & Chiang, H. H. (2020). Lightweight and resource-constrained learning network for face recognition with performance optimization. *Sensors (Switzerland)*, *20*(21), 1–20. https://doi.org/10.3390/s20216114

Post-training quantization | Google AI Edge. (2026, Mei 28). Google AI for Developers. https://ai.google.dev/edge/litert/conversion/tensorflow/quantization/post_training_quantization

Qualcomm Hexagon NPU. (t.t.). Diambil 9 Juni 2026, dari https://www.qualcomm.com/processors/hexagon

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2019). MobileNetV2: Inverted Residuals and Linear Bottlenecks. *arXiv:1801.04381 [cs.CV]*. http://arxiv.org/abs/1801.04381

Wang, M., & Deng, W. (2021). Deep face recognition: A survey. *Neurocomputing*, *429*, 215–244. https://doi.org/10.1016/j.neucom.2020.10.081

Xiao, J., Jiang, G., & Liu, H. (2022). A Lightweight Face Recognition Model based on MobileFaceNet for Limited Computation Environment. *EAI Endorsed Transactions on Internet of Things*, *7*(27), 1–9. https://doi.org/10.4108/eai.28-2-2022.173547