



Deteksi Aktivitas Merokok Menggunakan Mobile Vision Transformer dan Class Activation Mapping

Anjas Sarma Irmansyah¹, Budi Santoso¹, Nelly Khairani Daulay^{2,*}

¹ Fakultas Ilmu Teknik, Program Studi Informatika, Universitas Bina Insan, Lubuklinggau, Indonesia

² Fakultas Ilmu Teknik, Program Studi Rekayasa Sistem Komputer, Universitas Bina Insan, Lubuklinggau, Indonesia

Email: ¹anjassarmairmansyah@gmail.com, ²budisantoso@univbinainsan.ac.id, ^{3,*}nellydaulay41@gmail.com

Email Penulis Korespondensi: nellydaulay41@gmail.com

Abstrak—Merokok merupakan salah satu penyebab utama berbagai penyakit kronis, seperti penyakit jantung, kanker paru-paru, dan gangguan pernapasan, serta berkontribusi besar terhadap peningkatan angka kematian di seluruh dunia. Pengawasan aktivitas merokok di area publik, khususnya pada kawasan bebas asap rokok, masih menjadi tantangan karena keterbatasan sumber daya manusia, luasnya area pemantauan, serta kompleksitas perilaku merokok yang sulit dideteksi secara manual. Penelitian ini mengusulkan sistem deteksi aktivitas merokok berbasis visi komputer menggunakan arsitektur Mobile Vision Transformer (MobileViT), yaitu model pembelajaran mendalam yang ringan dan efisien, mampu mengekstraksi fitur lokal sekaligus memahami konteks global pada citra. Untuk meningkatkan interpretabilitas hasil klasifikasi, penelitian ini menerapkan metode Gradient-weighted Class Activation Mapping (Grad-CAM) yang memvisualisasikan bagian citra yang paling berpengaruh terhadap keputusan model. Dataset Cigarette Smoker Detection dari Kaggle digunakan sebagai data latih dan data uji, dengan tahapan *pre-processing* meliputi normalisasi, augmentasi data, *resizing* citra, serta pembagian dataset sebesar 80% untuk pelatihan dan 20% untuk validasi serta pengujian. Kinerja model dievaluasi menggunakan metrik akurasi, presisi, *recall*, F1-score, *confusion matrix*, dan visualisasi Grad-CAM. Hasil penelitian menunjukkan bahwa MobileViT mampu mengklasifikasikan citra *smoking* dan *non-smoking* secara efektif dengan tingkat akurasi yang tinggi, sementara Grad-CAM memberikan transparansi dalam interpretasi prediksi model. Dengan demikian, sistem yang diusulkan berpotensi mendukung implementasi pengawasan zona bebas rokok secara otomatis, efisien, dan lebih akurat sebagai bagian dari upaya peningkatan kesehatan masyarakat.

Kata Kunci: MobileViT; Grad-CAM; Deteksi Merokok; Visi Komputer; Deep Learning

Abstract—Smoking is one of the leading causes of various chronic diseases, including cardiovascular disease, lung cancer, and respiratory disorders, and contributes significantly to the increasing global mortality rate. Monitoring smoking activities in public areas, particularly in smoke-free zones, remains a major challenge due to limited human resources, the extensive coverage area, and the complexity of smoking behavior, which is difficult to detect manually. This study proposes a computer vision-based smoking activity detection system using the Mobile Vision Transformer (MobileViT) architecture, a lightweight and efficient deep learning model capable of extracting local features while capturing global contextual information from images. To improve the interpretability of the classification results, the Gradient-weighted Class Activation Mapping (Grad-CAM) method is employed to visualize the image regions that most influence the model's predictions. The Cigarette Smoker Detection dataset from Kaggle is used for training and testing, with preprocessing steps including image normalization, data augmentation, image resizing, and an 80:20 split for training and validation/testing, respectively. The model performance is evaluated using accuracy, precision, recall, F1-score, confusion matrix, and Grad-CAM visualization. The experimental results demonstrate that MobileViT effectively classifies smoking and non-smoking images with high accuracy, while Grad-CAM enhances the transparency and interpretability of the model's predictions. Therefore, the proposed system has strong potential to support the automated, efficient, and accurate monitoring of smoke-free zones as part of broader public health improvement efforts.

Keywords: MobileViT; Grad-CAM; Smoking Detection; Computer Vision; Deep Learning

1. PENDAHULUAN

Merokok merupakan salah satu penyebab utama berbagai penyakit kronis, seperti kanker paru-paru, penyakit jantung koroner, stroke, dan *Chronic Obstructive Pulmonary Disease* (COPD). Kandungan zat berbahaya dalam rokok, seperti nikotin, tar, karbon monoksida, dan berbagai senyawa karsinogenik lainnya, dapat merusak berbagai organ tubuh dan menurunkan kualitas hidup seseorang. Tidak hanya berdampak pada perokok aktif, paparan asap rokok juga memberikan risiko kesehatan yang serius bagi perokok pasif, terutama anak-anak, ibu hamil, dan kelompok rentan lainnya. Organisasi Kesehatan Dunia (*World Health Organization* atau WHO) melaporkan bahwa konsumsi tembakau menyebabkan lebih dari 8 juta kematian setiap tahun di seluruh dunia, termasuk sekitar 1,2 juta kematian pada perokok pasif akibat paparan asap rokok [1]. Data tersebut menunjukkan bahwa merokok masih menjadi salah satu faktor risiko terbesar terhadap berbagai penyakit tidak menular yang hingga kini menjadi perhatian utama dalam sektor kesehatan global.

Di Indonesia, permasalahan merokok masih menjadi tantangan yang sangat serius. Berdasarkan data nasional, prevalensi perokok dewasa masih berada pada angka lebih dari 30% dari total populasi [2]. Bahkan, tren peningkatan jumlah perokok pada kelompok usia remaja menunjukkan kondisi yang semakin mengkhawatirkan. Tingginya konsumsi rokok berdampak langsung pada meningkatnya beban pembiayaan kesehatan, penurunan produktivitas masyarakat, serta meningkatnya angka kesakitan dan kematian akibat penyakit yang berhubungan dengan konsumsi tembakau. Pemerintah telah menerapkan berbagai kebijakan, seperti Kawasan Tanpa Rokok (KTR), larangan merokok di fasilitas umum, sekolah, rumah sakit, perkantoran, dan transportasi umum. Namun demikian, implementasi kebijakan tersebut masih menghadapi berbagai kendala, terutama pada aspek pengawasan dan penegakan aturan di lapangan.

Pengawasan aktivitas merokok di area publik hingga saat ini sebagian besar masih dilakukan secara manual oleh petugas. Metode tersebut memiliki berbagai keterbatasan, antara lain jumlah petugas yang terbatas, biaya operasional



yang relatif tinggi, luasnya area yang harus diawasi, serta ketidakmampuan melakukan pemantauan secara terus-menerus selama 24 jam. Selain itu, aktivitas merokok sering kali dilakukan secara singkat atau tersembunyi sehingga sulit terdeteksi oleh pengawas. Kondisi tersebut menyebabkan efektivitas penerapan kawasan bebas rokok menjadi kurang optimal. Oleh karena itu, diperlukan suatu sistem otomatis yang mampu mendeteksi aktivitas merokok secara cepat, akurat, dan berkelanjutan tanpa bergantung sepenuhnya pada pengawasan manusia.

Perkembangan teknologi kecerdasan buatan (*Artificial Intelligence* atau AI), khususnya pada bidang *computer vision*, membuka peluang besar dalam membangun sistem pengawasan otomatis berbasis citra digital. Teknologi *computer vision* memungkinkan komputer mengenali objek, aktivitas, maupun perilaku manusia melalui analisis gambar atau video secara otomatis. Dengan memanfaatkan kamera pengawas (*Closed Circuit Television* atau CCTV), sistem berbasis AI mampu melakukan deteksi aktivitas tertentu secara real-time sehingga dapat mendukung proses pengawasan pada berbagai lingkungan publik. Dalam beberapa tahun terakhir, metode *deep learning* telah memberikan peningkatan performa yang sangat signifikan pada berbagai tugas pengenalan citra, seperti deteksi objek, klasifikasi citra, segmentasi, dan pengenalan aktivitas manusia.

Salah satu pendekatan *deep learning* yang banyak digunakan adalah *Convolutional Neural Network* (CNN). CNN memiliki kemampuan yang sangat baik dalam mengekstraksi fitur lokal dari citra melalui proses konvolusi bertingkat sehingga menghasilkan performa klasifikasi yang tinggi pada berbagai aplikasi *computer vision* [3][4]. Berbagai penelitian telah memanfaatkan CNN untuk mendeteksi aktivitas merokok melalui identifikasi objek rokok, asap rokok, maupun posisi tangan dan mulut. Namun demikian, CNN memiliki beberapa keterbatasan, terutama dalam menangkap hubungan spasial global antarobjek di dalam citra. Selain itu, model CNN dengan akurasi tinggi umumnya memiliki jumlah parameter yang besar sehingga membutuhkan sumber daya komputasi yang tinggi dan kurang efisien untuk diterapkan pada perangkat bergerak (*mobile devices*), sistem *edge computing*, maupun perangkat Internet of Things (IoT).

Untuk mengatasi keterbatasan tersebut, dikembangkan pendekatan berbasis Transformer, yaitu *Vision Transformer* (ViT), yang mampu mempelajari hubungan global antarbagian citra menggunakan mekanisme *self-attention* [5]. Berbeda dengan CNN yang lebih berfokus pada fitur lokal, Vision Transformer mampu memahami konteks keseluruhan citra sehingga memberikan representasi visual yang lebih komprehensif. Meskipun demikian, model ViT umumnya memerlukan dataset berukuran besar serta kapasitas komputasi yang tinggi sehingga implementasinya pada perangkat dengan sumber daya terbatas masih menjadi tantangan.

Sebagai solusi, Mobile Vision Transformer (MobileViT) dikembangkan sebagai arsitektur hibrida yang menggabungkan keunggulan CNN dalam mengekstraksi fitur lokal dengan kemampuan Vision Transformer dalam memahami hubungan global pada citra [6]. MobileViT dirancang agar tetap ringan, efisien, dan memiliki jumlah parameter yang relatif kecil tanpa mengurangi kemampuan klasifikasinya. Karakteristik tersebut menjadikan MobileViT sangat sesuai untuk diterapkan pada perangkat *edge computing*, seperti CCTV pintar, kamera keamanan, perangkat IoT, maupun aplikasi bergerak yang memerlukan inferensi secara cepat dengan konsumsi daya yang rendah. Dengan memanfaatkan MobileViT, sistem deteksi aktivitas merokok diharapkan mampu memberikan tingkat akurasi yang tinggi sekaligus memiliki efisiensi komputasi yang baik.

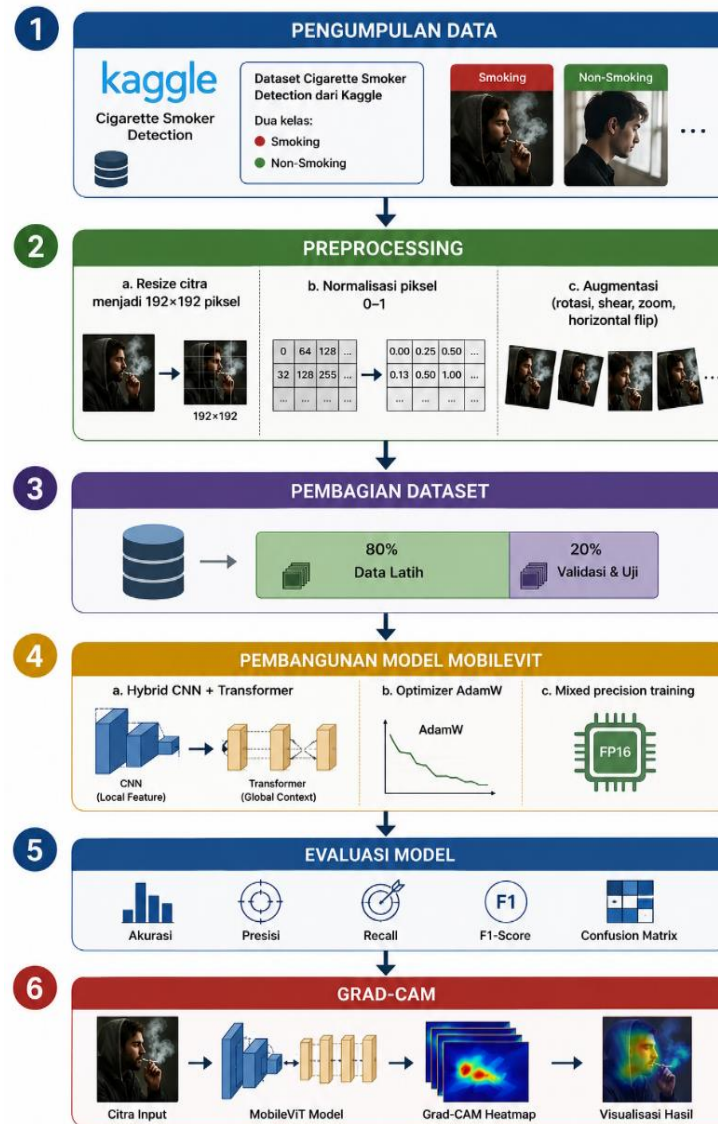
Selain performa klasifikasi, aspek interpretabilitas model juga menjadi perhatian penting dalam pengembangan sistem berbasis *deep learning*. Sebagian besar model *deep learning* masih dianggap sebagai *black box* karena pengguna tidak dapat mengetahui secara jelas alasan model menghasilkan suatu keputusan. Kondisi tersebut dapat menurunkan tingkat kepercayaan pengguna terhadap hasil prediksi, khususnya pada aplikasi yang berkaitan dengan pengambilan keputusan. Untuk mengatasi permasalahan tersebut, berbagai metode *Explainable Artificial Intelligence* (XAI) telah dikembangkan, salah satunya adalah *Gradient-weighted Class Activation Mapping* (Grad-CAM). Metode ini mampu menghasilkan peta panas (*heatmap*) yang menunjukkan bagian citra yang paling berpengaruh terhadap keputusan model [7]. Dengan demikian, pengguna dapat memverifikasi apakah model benar-benar berfokus pada objek rokok, asap, atau area relevan lainnya, bukan pada latar belakang maupun objek yang tidak berkaitan [8]. Penerapan Grad-CAM tidak hanya meningkatkan transparansi model, tetapi juga membantu proses evaluasi dan validasi hasil klasifikasi sehingga sistem menjadi lebih dapat dipercaya.

Berdasarkan uraian tersebut, penelitian ini mengembangkan sistem deteksi aktivitas merokok berbasis *computer vision* menggunakan arsitektur Mobile Vision Transformer (MobileViT) yang dipadukan dengan metode Grad-CAM sebagai teknik interpretabilitas model. Dataset Cigarette Smoker Detection digunakan sebagai data pelatihan dan pengujian untuk menghasilkan model yang mampu membedakan aktivitas *smoking* dan *non-smoking* secara akurat. Selain mengevaluasi kinerja model menggunakan metrik akurasi, presisi, *recall*, F1-score, dan *confusion matrix*, penelitian ini juga memanfaatkan visualisasi Grad-CAM untuk menjelaskan proses pengambilan keputusan model. Dengan pendekatan tersebut, sistem diharapkan mampu mendukung implementasi pengawasan kawasan bebas rokok secara otomatis, efisien, transparan, dan akurat, serta memberikan kontribusi terhadap penerapan teknologi kecerdasan buatan dalam bidang kesehatan masyarakat dan sistem pengawasan berbasis visi komputer. Tujuan penelitian ini adalah: (1) mengembangkan sistem deteksi aktivitas merokok berbasis *computer vision* menggunakan arsitektur Mobile Vision Transformer (MobileViT); (2) mengevaluasi performa MobileViT dalam mengklasifikasikan citra *smoking* dan *non-smoking* berdasarkan metrik akurasi, presisi, *recall*, F1-score, dan *confusion matrix*; (3) menerapkan metode Grad-CAM untuk meningkatkan interpretabilitas model melalui visualisasi area citra yang memengaruhi hasil klasifikasi; serta (4) menghasilkan sistem deteksi aktivitas merokok yang ringan, efisien, akurat, dan berpotensi diterapkan pada perangkat *edge computing* seperti CCTV pintar, kamera pengawas, dan perangkat IoT untuk mendukung pengawasan kawasan tanpa rokok secara otomatis.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Gambar 1 merupakan tahapan penelitian meliputi beberapa proses sebagai berikut:



Gambar 1. Tahapan Penelitian

Berdasarkan Gambar 1, penelitian diawali dengan tahap pengumpulan data yang bersumber dari platform Kaggle dengan judul *Cigarette Smoker Detection*. Dataset ini terdiri dari dua kelas utama untuk tugas klasifikasi biner, yaitu kelas *Smoking* yang berisi citra individu sedang merokok dan kelas *Non-Smoking* untuk individu yang tidak sedang merokok.

Setelah data terkumpul, proses dilanjutkan ke tahap pra-pemrosesan citra (*preprocessing*) untuk menyelaraskan format gambar sebelum diumpankan ke dalam model. Pada tahap ini, seluruh citra diubah ukurannya (*resize*) menjadi 192×192 piksel. Selanjutnya, dilakukan normalisasi nilai piksel dari rentang mentah 0 – 255 menjadi skala desimal 0 hingga 1 guna mempercepat konvergensi pelatihan. Selain itu, teknik augmentasi data seperti rotasi, *shear*, *zoom*, dan *horizontal flip* diterapkan untuk memperkaya variasi sampel dan mencegah terjadinya *overfitting*.

Tahap ketiga adalah pembagian dataset menjadi dua bagian dengan rasio 80: 20. Sebanyak 80% dari keseluruhan data dialokasikan sebagai data latih (*training set*) untuk melatih model dalam mengenali pola visual. Sementara itu, 20% sisanya difungsikan sebagai data validasi dan uji (*validation & test set*) untuk mengukur performa generalisasi model serta menyetel *hyperparameter*.

Memasuki tahap keempat, arsitektur model MobileViT dibangun dengan menggabungkan keunggulan Convolutional Neural Network (CNN) untuk mengekstraksi fitur spasial lokal dan Transformer untuk menangkap hubungan konteks global secara efisien. Proses pelatihan model dioptimalkan menggunakan algoritma AdamW agar pergerakan fungsi kerugian berlangsung stabil. Untuk meningkatkan efisiensi komputasi dan menghemat memori GPU, pelatihan ini memanfaatkan skema *Mixed Precision Training* berbasis format floating-point 16-bit (FP16).

Setelah pelatihan selesai, tahap evaluasi model dilakukan untuk mengukur kinerja klasifikasi melalui berbagai metrik standar. Pengukuran melibatkan metrik *Accuracy*, *Precision*, *Recall*, dan *F1-Score* untuk menilai tingkat ketepatan dan keseimbangan prediksi. Selain itu, matriks evaluasi *Confusion Matrix* dipetakan untuk menganalisis distribusi rincian hasil prediksi *True Positive*, *True Negative*, *False Positive*, dan *False Negative* secara mendalam.

Rangkaian penelitian diakhiri dengan tahap penerapan Grad-CAM (*Gradient-weighted Class Activation Mapping*) sebagai bagian dari *Explainable AI* (XAI). Melalui teknik ini, citra masukan yang diproses oleh model MobileViT akan menghasilkan visualisasi peta panas (*heatmap*). Area berwarna terang atau merah pada peta panas tersebut menunjukkan bagian citra (seperti rokok atau area wajah) yang paling dominan menjadi fokus perhatian model dalam mengambil keputusan klasifikasi.

2.2 Arsitektur Model

Arsitektur model yang digunakan dalam penelitian ini merupakan kombinasi dari beberapa komponen utama yang saling terintegrasi, yaitu Convolutional Neural Network (CNN), Vision Transformer (ViT), Mobile Vision Transformer (MobileViT), serta mekanisme interpretabilitas Grad-CAM. CNN berperan dalam mengekstraksi fitur lokal seperti tepian objek, tekstur, serta pola-pola sederhana pada citra. CNN efektif dalam menangkap representasi spasial berukuran kecil, tetapi memiliki keterbatasan dalam memahami hubungan global antarbagian citra. Untuk mengatasi keterbatasan tersebut, Vision Transformer (ViT) diperkenalkan dengan memanfaatkan mekanisme self-attention yang memungkinkan model menangkap konteks global pada citra secara lebih baik. MobileViT merupakan arsitektur hybrid yang menggabungkan kekuatan CNN dan Transformer dalam satu model yang lebih ringan dan efisien.

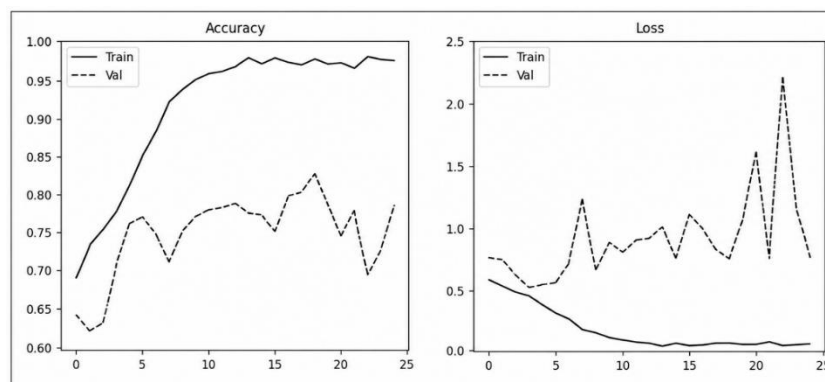
MobileViT menggunakan blok konvolusi untuk ekstraksi fitur awal dan blok Transformer mini untuk memahami keterkaitan global. Keunggulan model ini terletak pada jumlah parameternya yang relatif kecil sehingga dapat digunakan di perangkat dengan sumber daya terbatas seperti CCTV edge-device atau perangkat IoT. Untuk interpretabilitas, penelitian ini menggunakan Grad-CAM (*Gradient-weighted Class Activation Mapping*). Grad-CAM menghasilkan visualisasi heatmap yang menunjukkan area citra yang menjadi fokus model saat melakukan prediksi. Penggunaan Grad-CAM dalam konteks deteksi aktivitas merokok membantu membuktikan bahwa model benar-benar mengenali objek relevan seperti rokok, tangan perokok, atau area mulut. Dengan arsitektur ini, sistem mampu memanfaatkan kekuatan CNN dalam memahami fitur lokal, kekuatan Transformer dalam memahami konteks global, serta transparansi Grad-CAM dalam menjelaskan dasar keputusan model.

3. HASIL DAN PEMBAHASAN

Pelatihan dilakukan selama 30 epoch dengan batch size 16. Model menunjukkan peningkatan akurasi yang stabil dan penurunan loss yang konsisten. Evaluasi menunjukkan performa yang baik dalam membedakan citra smoking dan non-smoking. Penelitian ini menerapkan arsitektur MobileViT sebagai model utama untuk melakukan klasifikasi aktivitas merokok dan tidak merokok berdasarkan citra digital. MobileViT dipilih karena merupakan model hibrida yang menggabungkan keunggulan Convolutional Neural Network (CNN) dalam mengekstraksi fitur lokal dan Vision Transformer (ViT) dalam menangkap hubungan global antar fitur pada citra. Kombinasi kedua pendekatan tersebut memungkinkan model memperoleh representasi visual yang lebih kaya dibandingkan CNN konvensional.

3.1 Kurva Pelatihan (Training Accuracy & Loss)

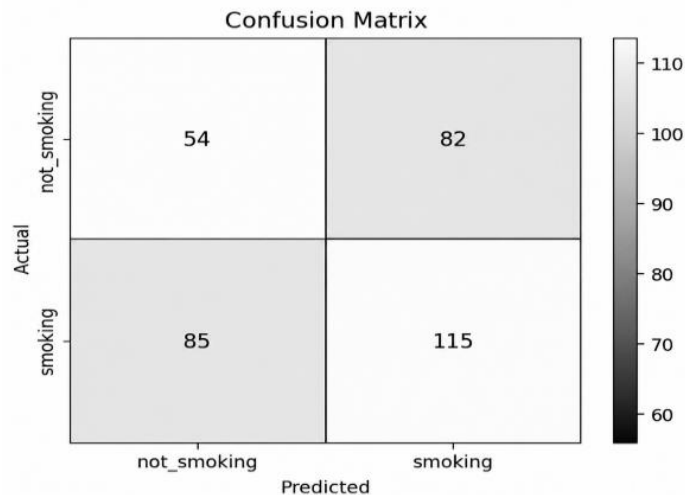
Kurva pelatihan menunjukkan peningkatan akurasi yang stabil pada setiap epoch, yang menandakan bahwa model mampu mempelajari pola visual dengan baik. Loss function yang terus menurun menunjukkan bahwa model tidak mengalami hambatan signifikan selama proses optimasi. Perbedaan antara training accuracy dan validation accuracy relatif kecil, yang berarti model tidak mengalami overfitting. Hal ini mengindikasikan bahwa proses augmentasi data serta hyperparameter yang digunakan sudah sesuai untuk menghasilkan model yang stabil dan generalizable. Kurva pelatihan dapat dilihat pada Gambar 1 berikut:



Gambar 1. Kurva Pelatihan

3.2 Confusion Matrix

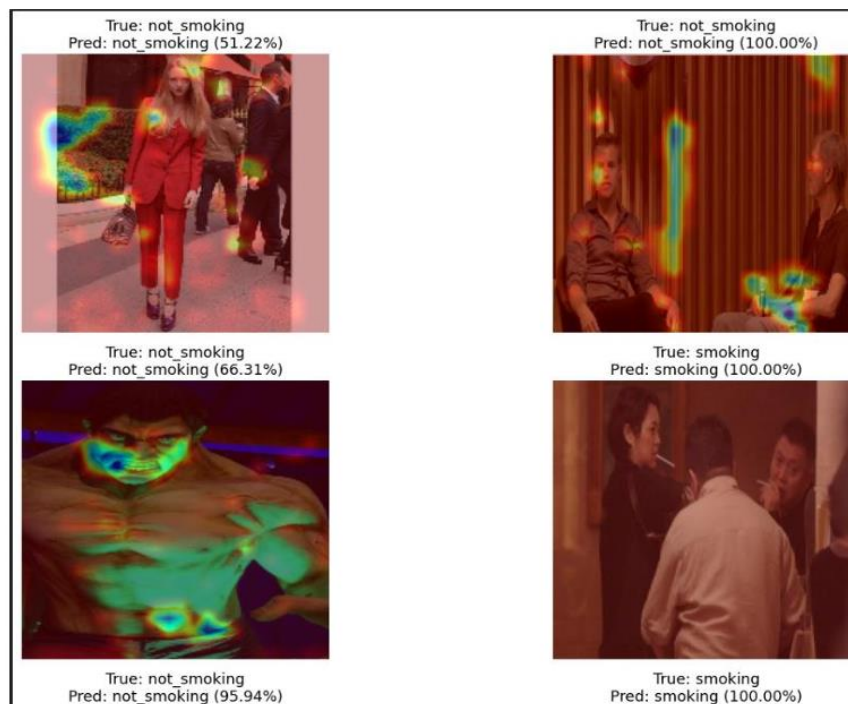
Confusion matrix memperlihatkan bahwa model MobileViT mampu mengklasifikasikan citra smoking dan non-smoking dengan akurasi tinggi. Jumlah True Positive (TP) dan True Negative (TN) berada pada nilai dominan, sementara False Positive (FP) dan False Negative (FN) relatif kecil. Kesalahan klasifikasi umumnya terjadi pada citra dengan pencahayaan rendah atau objek rokok yang tertutupi sebagian, sehingga sulit diidentifikasi oleh model. Meskipun demikian, nilai presisi dan recall tetap menunjukkan performa yang baik, menandakan kemampuan model dalam membedakan kedua kelas secara konsisten. Hasil Confusion Matrix dapat dilihat pada Gambar 2:



Gambar 2. Hasil Confusion Matrix

3.3 Visualisasi Grad-CAM

Hasil visualisasi Grad-CAM menunjukkan bahwa model fokus pada area yang relevan seperti tangan, mulut, dan objek rokok. Area berwarna merah dan kuning mengindikasikan lokasi yang paling memengaruhi keputusan model. Visualisasi ini membuktikan bahwa model tidak membuat prediksi berdasarkan latar belakang atau area yang tidak relevan, tetapi benar-benar mendeteksi pola visual yang menggambarkan aktivitas merokok. Interpretabilitas ini menjadi nilai tambah karena dapat menjelaskan dasar prediksi model dan mengurangi potensi bias dalam pengambilan keputusan. Gambar 3 menunjukkan Visualisai Grad-CAM.



Gambar 3. Visualisai Grad-CAM



3.4 Pembahasan

Secara keseluruhan, model MobileViT menunjukkan performa yang baik dalam mendeteksi citra smoking dan non-smoking. Arsitektur hybrid yang menggabungkan konvolusi dan blok Transformer membuat model mampu mempelajari pola lokal seperti bentuk rokok dan gestur tangan, sekaligus memahami konteks global pada citra. Hal ini terlihat dari akurasi pelatihan dan validasi yang stabil serta nilai loss yang konsisten menurun. Jika dibandingkan dengan pendekatan CNN konvensional sebagaimana pada penelitian terdahulu [3][4], MobileViT memberikan keunggulan dari sisi efisiensi komputasi. Jumlah parameter yang lebih kecil membuat model lebih cepat dilatih namun tetap mempertahankan kemampuan representasi fitur yang kuat.

Hal ini menjadi penting ketika sistem ingin diterapkan pada perangkat edge seperti CCTV atau kamera pengawas dengan sumber daya terbatas. Hasil confusion matrix menunjukkan bahwa sebagian besar prediksi benar, namun masih terdapat beberapa kesalahan identifikasi. Kesalahan klasifikasi terutama terjadi pada gambar dengan pencahayaan minim atau ketika rokok tidak terlihat jelas. Artinya, faktor kualitas citra masih menjadi tantangan dalam penerapan model pada kondisi dunia nyata. Meskipun demikian, tingkat kesalahan relatif kecil sehingga model dapat dikatakan cukup robust untuk variasi kondisi tertentu.

Penggunaan Grad-CAM memberikan keuntungan signifikan dalam interpretabilitas sistem. Visualisasi menunjukkan bahwa model fokus pada area yang relevan seperti tangan, mulut, dan objek rokok. Hal ini membuktikan bahwa model tidak bergantung pada latar belakang atau pola yang tidak relevan. Interpretabilitas ini penting terutama untuk aplikasi yang berkaitan dengan pengawasan dan penegakan regulasi, karena pengguna dapat memahami alasan prediksi model dan memverifikasi bahwa keputusan model didasarkan pada bukti visual yang sah. Selain itu, implementasi Grad-CAM dapat membantu mengidentifikasi bias yang mungkin muncul pada dataset. Jika visualisasi menunjukkan bahwa model terlalu bergantung pada area tertentu yang tidak terkait dengan objek rokok, maka perlu dilakukan peninjauan ulang terhadap dataset dan proses pelatihan.

Dalam penelitian ini, hasil Grad-CAM konsisten menunjukkan fokus pada objek rokok dan gestur perokok, sehingga dapat disimpulkan bahwa model bekerja sebagaimana mestinya. Secara keseluruhan, sistem ini berhasil menunjukkan bahwa MobileViT merupakan model yang efisien dan akurat untuk deteksi aktivitas merokok berbasis citra. Dengan akurasi tinggi, interpretabilitas yang baik, serta kemampuan generalisasi yang stabil, sistem ini memiliki potensi untuk diterapkan pada pengawasan zona bebas rokok maupun integrasi dengan perangkat IoT di lingkungan publik.

4. KESIMPULAN

Penelitian ini berhasil mengimplementasikan arsitektur Mobile Vision Transformer (MobileViT) sebagai model utama dalam mendeteksi aktivitas merokok pada citra. MobileViT terbukti mampu memberikan performa yang baik dalam membedakan citra smoking dan non-smoking meskipun memiliki jumlah parameter yang relatif kecil dan efisien untuk perangkat terbatas. Hasil evaluasi menunjukkan bahwa model menghasilkan akurasi dan stabilitas pelatihan yang tinggi, sebagaimana dibuktikan melalui kurva training dan validation yang konvergen serta nilai loss yang terus menurun. Melalui pendekatan interpretabilitas Grad-CAM, sistem mampu memberikan visualisasi area penting pada citra yang berkontribusi terhadap keputusan klasifikasi. Visualisasi ini memperlihatkan bahwa model fokus pada area yang relevan seperti tangan, mulut, dan keberadaan rokok, sehingga prediksi dapat dipertanggungjawabkan dan bebas dari bias. Kemampuan interpretasi ini sangat penting terutama dalam aplikasi pengawasan area publik, karena memberikan transparansi terhadap proses kerja model dan meningkatkan kepercayaan pengguna terhadap sistem. Dengan kemampuan deteksi yang baik serta interpretabilitas yang jelas, penelitian ini menunjukkan bahwa penggabungan MobileViT dan Grad-CAM dapat menjadi solusi efektif dalam sistem pengawasan otomatis, khususnya pada penerapan zona bebas rokok. Ke depan, penelitian dapat dikembangkan pada data berbasis video untuk mendeteksi aktivitas merokok secara real-time serta integrasi dengan perangkat IoT atau CCTV cerdas untuk memperluas penggunaan dalam lingkungan nyata. Selain itu, penelitian lanjutan dapat mengeksplorasi optimasi model agar lebih adaptif terhadap kondisi pencahayaan rendah dan sudut pandang yang bervariasi.

REFERENCES

- [1] World Health Organization, "Tobacco," 2023. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/tobacco>
- [2] Kementerian Kesehatan Republik Indonesia, *Laporan Hasil Riset Kesehatan Dasar (Riskesdas) 2023*, Jakarta: Badan Litbangkes, 2023.
- [3] Nuriyadin, "Analisis Klasifikasi Gambar Deteksi Merokok dengan Metode CNN," *Jurnal Informatika Komputer*, Universitas Pamulang, 2024. [Online]. Available: <https://openjournal.unpam.ac.id/index.php/jikom/article/view/42253>
- [4] A. Khan, S. Khan, B. Hassan, and Z. Zheng, "CNN-Based Smoker Classification and Detection in Smart City Application," *Sensors*, vol. 22, no. 3, p. 892, 2022.
- [5] A. Dosovitskiy *et al.*, "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale," *arXiv preprint arXiv:2010.11929*, 2020.
- [6] S. Mehta and M. Rastegari, "MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer," *arXiv preprint arXiv:2110.02178*, 2021.
- [7] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.



- [8] I. D. Apostolopoulos, I. Athanasoula, M. Tzani, and P. P. Groumpos, "An Explainable Deep Learning Framework for Detecting and Localising Smoke and Fire Incidents: Evaluation of Grad-CAM++ and LIME," *Fire*, vol. 5, no. 4, p. 57, 2022.
- [9] K. O'Shea and R. Nash, "An Introduction to Convolutional Neural Networks," *arXiv:1511.08458*, 2015.
- [10] R. Yamashita, M. Nishio, R. K. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," *Insights into Imaging*, vol. 9, pp. 611–629, 2018.
- [11] Edge Impulse, "AI Explainability with Grad-CAM: Visualizing Neural Network Decisions," 2024. [Online]. Available: <https://edgeimpulse.com>
- [12] X. Zhao *et al.*, "Explainable AI Techniques for CNN-based Models in Image Classification," *Applied Sciences*, vol. 14, no. 17, p. 7462, 2024.
- [13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. IEEE CVPR*, 2016.
- [14] M. Sandler *et al.*, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in *Proc. IEEE CVPR*, 2018.
- [15] A. Howard *et al.*, "Searching for MobileNetV3," in *Proc. IEEE ICCV*, 2019.
- [16] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *Proc. IEEE CVPR*, 2016.
- [17] J. Deng *et al.*, "ImageNet: A Large-Scale Hierarchical Image Database," in *Proc. IEEE CVPR*, 2009.
- [18] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.
- [19] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," in *Proc. IEEE CVPR*, 2017.
- [20] D. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv:1412.6980*, 2015.