

ANALISIS *EXPLAINABILITY DEEP LEARNING* UNTUK KLASIFIKASI TUMOR OTAK PADA MRI T2-WEIGHTED MENGUNAKAN *GRADIENT-WEIGHTED CLASS ACTIVATION MAPPING (GRAD-CAM)*

¹Davin Valerian, ²Windra Swastika¹

^{1,2}*Program Studi Teknik Informatika, Fakultas Teknologi dan Desain, Universitas Ma Chung, Villa Puncak
Tidar N-01, Malang, Jawa Timur, Indonesia, 65151*

***Email korespondensi:** 312310012@student.machung.ac.id

Received: 27 05 2026 – Revised: 29 07 2026 - Accepted: 04 08 2026 - Published: 05 08 2026

Abstrak. Tumor otak merupakan kondisi medis kritis yang memerlukan deteksi dini secara akurat, dengan Magnetic Resonance Imaging (MRI) T2-Weighted sebagai modalitas standar emas. Meskipun model deep learning berbasis Convolutional Neural Network (CNN) mampu mengklasifikasikan tumor otak dengan akurasi tinggi, sifatnya yang black box menghambat kepercayaan tenaga medis terhadap dasar pengambilan keputusan model. Penelitian ini bertujuan membangun sistem klasifikasi tumor otak berbasis EfficientNetB0 dengan pendekatan transfer learning, serta menerapkan Gradient-weighted Class Activation Mapping (Grad-CAM) secara mandiri melalui mekanisme tf.GradientTape untuk mengungkap area citra yang mendasari keputusan model. Dataset yang digunakan adalah Brain Tumor Classification (MRI) dari Kaggle sejumlah 3.264 citra T2-Weighted dengan empat kelas, yaitu glioma, meningioma, pituitary, dan no tumor. Strategi pelatihan dua fase, yaitu feature extraction dan fine-tuning selektif pada 20 lapisan terakhir backbone, diterapkan untuk mengoptimalkan performa klasifikasi. Hasil pengujian menunjukkan model final mencapai F1-macro sebesar 0,8595 pada test set, melampaui target 0,85 dan meningkat signifikan dari model baseline (0,6938). Evaluasi Grad-CAM terhadap 80 sampel stratified melalui Expert Simple Assessment menunjukkan 57,5% heatmap dinilai baik secara anatomis, namun mengungkap kelemahan sistemik pada kelas glioma (F1-score 0,6839; skor explainability rata-rata 0,90 dari 2) yang disertai anomali confidence score yang terbalik serta sebelas kasus high-confidence wrong. Temuan ini menegaskan bahwa evaluasi akurasi makro semata tidak cukup untuk menjamin keandalan klinis model deep learning, sehingga explainability perlu menjadi bagian integral dari proses validasi sistem Computer-Aided Detection sebelum diadopsi dalam praktik medis.

Kata kunci: *Explainable AI, Grad-CAM, EfficientNetB0, Klasifikasi Tumor Otak, MRI T2-Weighted*

PENDAHULUAN

Laboratorium Artificial Intelligence for Digital Images and Technopreneurship (AiDiTech) Universitas Ma Chung merupakan laboratorium riset yang berfokus pada kecerdasan buatan, khususnya machine learning dan deep learning untuk identifikasi citra medis, dengan rekam jejak pada deteksi parasit malaria dan Diabetic Retinopathy. Sejalan dengan fokus riset tersebut, penelitian ini mengangkat tema klasifikasi citra medis tingkat lanjut pada kasus tumor otak.

Tumor otak merupakan kondisi medis kritis yang memerlukan deteksi dini secara akurat untuk meningkatkan angka kelangsungan hidup pasien. Standar emas dalam mendiagnosis kondisi ini adalah penggunaan modalitas Magnetic Resonance Imaging (MRI) T2-Weighted. Akan tetapi, interpretasi visual terhadap hasil pemindaian MRI secara manual membutuhkan waktu, keahlian radiolog yang sangat spesifik, bersifat subjektif, serta rentan terhadap human error.

Sistem Computer-Aided Detection (CAD) berbasis deep learning, khususnya Convolutional Neural Network (CNN) dengan metode transfer learning, telah terbukti mampu memberikan

tingkat akurasi tinggi dalam mengklasifikasikan jenis tumor. Meskipun demikian, model deep learning medis sering kali terhambat dalam tahap implementasi klinis karena sifatnya yang black box, di mana proses dan dasar keputusan model sulit dipahami oleh tenaga medis. Untuk memecahkan masalah kepercayaan dan transparansi tersebut, diperlukan pendekatan Explainable AI (XAI). Penelitian ini mengusulkan penerapan Gradient-weighted Class Activation Mapping (Grad-CAM) untuk mendampingi model klasifikasi EfficientNetB0, karena kemampuannya menghasilkan heatmap visual yang melokalisasi area prediksi model secara anatomis, sehingga radiolog dapat memvalidasi apakah model mendeteksi tumor pada wilayah yang relevan secara klinis.

Penelitian ini bertujuan (1) membangun dan mengoptimasi pipeline klasifikasi citra medis multi-kelas berbasis EfficientNetB0 dengan target F1-macro minimal 0,85 pada test set; (2) mengimplementasikan algoritma Grad-CAM secara mandiri untuk menghasilkan heatmap localization; (3) mengevaluasi plausibilitas anatomis heatmap secara kualitatif dan kuantitatif; serta (4) membedah karakteristik kegagalan model melalui klasifikasi tipe misprediksi untuk memetakan potensi risiko klinis.

MASALAH

Persoalan utama yang diangkat dalam penelitian ini adalah rendahnya transparansi model deep learning dalam mengklasifikasikan tumor otak pada citra MRI T2-Weighted. Model CNN berbasis transfer learning umumnya mampu mencapai akurasi klasifikasi yang tinggi, namun tidak memberikan penjelasan mengenai dasar spasial dari keputusan yang diambil, sehingga sulit divalidasi oleh tenaga medis sebelum digunakan sebagai alat bantu diagnostik. Ruang lingkup penelitian dibatasi pada klasifikasi empat kategori, yaitu glioma tumor, meningioma tumor, pituitary tumor, dan no tumor, tanpa mencakup segmentasi otomatis atau lokalisasi koordinat bounding box karena ketiadaan anotasi koordinat pada dataset yang digunakan. Model dibangun menggunakan arsitektur EfficientNetB0 dengan bobot pre-trained ImageNet, dioptimasi melalui fine-tuning pada 20 lapisan terakhir backbone, sedangkan aspek explainability diekstraksi secara manual melalui implementasi Grad-CAM menggunakan operasi `tf.GradientTape`.

METODE PELAKSANAAN

Penelitian ini menggunakan pendekatan simulasi dan penerapan ilmu pengetahuan dan teknologi (Ipteks) melalui pembangunan model deep learning serta teknik XAI untuk menjelaskan dan memvalidasi hasil klasifikasi. Dataset yang digunakan adalah Brain Tumor Classification (MRI) yang dikompilasi oleh Sartaj Bhuvaji dan tersedia secara publik di platform Kaggle, terdiri dari 3.264 citra MRI T2-Weighted dengan distribusi sebagaimana disajikan pada Tabel 1.

Tabel 1. Distribusi Dataset per Kelas

Kelas	Training	Testing	Total
glioma_tumor	826	100	926
meningioma_tumor	822	115	937
no_tumor	395	105	500
pituitary_tumor	827	74	901
TOTAL	2.870	394	3.264

Ditemukan class imbalance dengan rasio 2,09x antara kelas mayoritas (pituitary_tumor) dan minoritas (no_tumor), sehingga diterapkan class weighting secara inverse-proportional pada fungsi loss. Pipeline praproses citra dirancang menjadi tiga tahap sekuensial: (1) Brain Region Cropping menggunakan Gaussian Blur, binary thresholding, operasi morfologi erosi-dilasi, dan ekstraksi kontur maksimal untuk memotong citra tepat pada batas otak; (2) CLAHE (Contrast Limited Adaptive Histogram Equalization) pada channel L ruang warna LAB dengan clipLimit=2,0 dan tileGridSize=(8x8) untuk meningkatkan kontras lokal tanpa amplifikasi noise berlebih; serta (3) resize ke 224x224 piksel dan konversi ke float32 sesuai kebutuhan input EfficientNetB0.

Model dibangun menggunakan Keras Functional API dengan EfficientNetB0 sebagai backbone feature extractor (238 lapisan, ± 4 juta parameter) yang dilengkapi classification head berupa GlobalAveragePooling2D, Dense(256, ReLU), Dropout(0,3), dan Dense(4, Softmax). Strategi pelatihan dirancang dalam dua fase: Fase 1 (baseline/feature extraction) dengan backbone dibekukan sepenuhnya, optimizer Adam learning rate $1e-4$, maksimum 20 epoch, dan Early Stopping (patience=5); serta Fase 2 (fine-tuning selektif) yang membuka 20 lapisan terakhir backbone dengan learning rate diturunkan menjadi $5e-5$ disertai callback ReduceLROnPlateau untuk mencegah catastrophic forgetting.

Implementasi Grad-CAM menggunakan pendekatan manual forward pass per layer melalui mekanisme tf.GradientTape, dipilih karena arsitektur nested model pada Keras 3.x tidak mendukung akses tensor lintas sub-model menggunakan metode Grad-CAM standar. Target layer yang digunakan adalah top_activation (feature map $7 \times 7 \times 1280$). Heatmap yang dihasilkan diperbesar ke 224x224 melalui upsampling bilinear, diwarnai dengan colormap JET, dan di-overlay pada citra asli dengan transparansi $\alpha=0,45$. Evaluasi explainability dilakukan pada 80 sampel stratified dari test set (20 citra per kelas) melalui dua pendekatan: (1) Expert Simple Assessment, penilaian kualitatif berskala 0–2 berdasarkan plausibilitas anatomis heatmap terhadap lokasi tumor yang diharapkan pada tiap kelas; dan (2) empat metrik kuantitatif otomatis, yaitu Heatmap Concentration Score (HCS), Foreground Ratio (FR), Activation Entropy (AE), dan Spatial Compactness (SC). Sebagai pembanding tambahan, dilakukan pula eksperimen Test-Time Augmentation (TTA) dengan lima pass augmentasi acak per citra. Evaluasi performa klasifikasi menggunakan metrik F1-macro, Precision, Recall, dan Confusion Matrix pada test set sejumlah 394 citra yang disisihkan sejak awal dan hanya digunakan satu kali di akhir eksperimen.

Seluruh implementasi menggunakan Python 3.10, TensorFlow 2.19, dan Keras 3.10, dengan OpenCV untuk pemrosesan citra serta NumPy, Pandas, Matplotlib, dan Scikit-learn untuk komputasi numerik, tabulasi, dan evaluasi. Eksperimen dijalankan pada lingkungan komputasi awan Kaggle Notebook dengan akselerasi dua unit GPU NVIDIA Tesla T4.

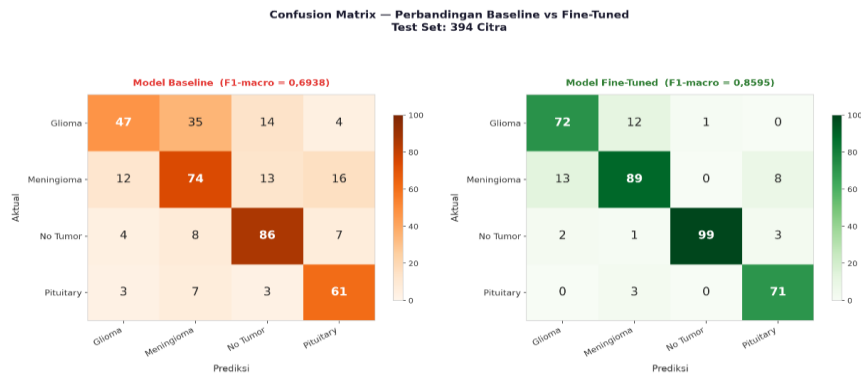
HASIL DAN PEMBAHASAN

Performa Klasifikasi: Baseline vs Fine-Tuned

Pengujian pada test set (394 citra) menunjukkan model baseline gagal memenuhi target KPI dengan F1-macro hanya 0,6938, mengonfirmasi bahwa pendekatan feature extraction murni tidak cukup mengadaptasikan fitur ImageNet ke domain citra MRI tumor otak. Strategi fine-tuning selektif terbukti esensial, meningkatkan F1-macro sebesar +0,1657 hingga mencapai 0,8595, sehingga target KPI F1-macro $\geq 0,85$ berhasil terpenuhi (Tabel 2).

Tabel 2. Perbandingan Performa Model Baseline vs Fine-Tuned

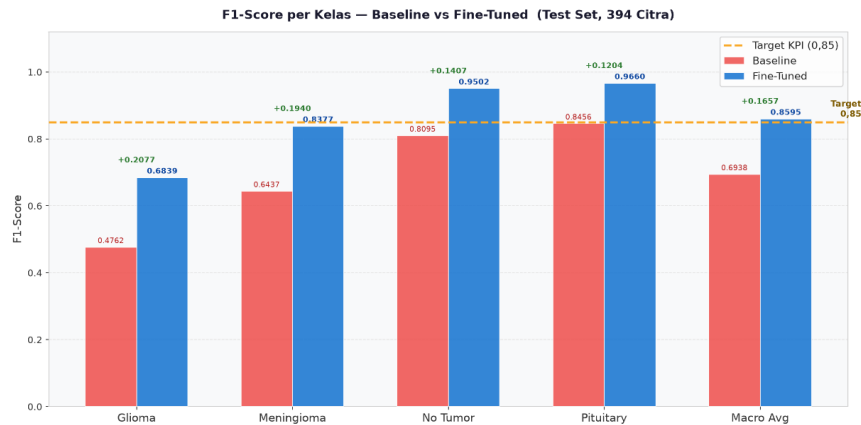
Metrik	Baseline	Fine-Tuned	Δ
Test Accuracy	69,38%	86,29%	+16,91%
Precision (Macro)	0,7033	0,8954	+0,1921
Recall (Macro)	0,6978	0,8637	+0,1659
F1-Macro	0,6938	0,8595	+0,1657



Gambar 1. Confusion Matrix Model Baseline (kiri) vs Fine-Tuned (kanan)

Analisis Performa per Kelas

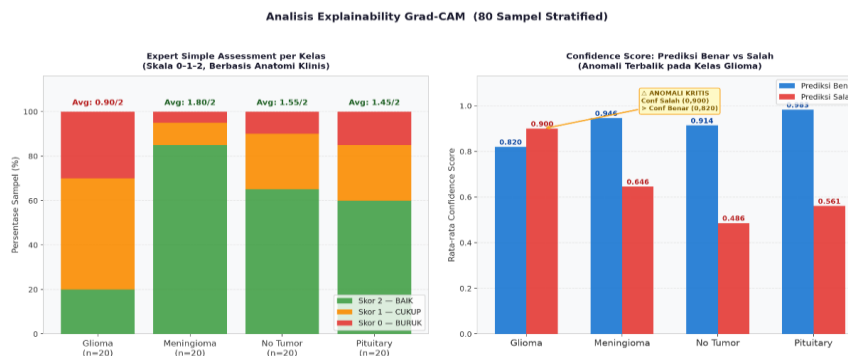
Meski secara makro target tercapai, analisis per kelas mengungkap distribusi performa yang tidak merata. Kelas no_tumor (F1=0,9502) dan pituitary_tumor (F1=0,9660) melampaui ambang 0,85 secara ketat, kelas meningioma_tumor (F1=0,8377) mendekati ambang, sedangkan kelas glioma_tumor dengan F1-score hanya 0,6839 menjadi titik kelemahan sistemik paling menonjol. Kelemahan ini konsisten dengan sifat biologis glioma yang dapat muncul di berbagai lokasi parenkim otak, sehingga variasinya sangat tinggi dibandingkan kelas lain.



Gambar 2. Perbandingan F1-Score per Kelas: Baseline vs Fine-Tuned

Evaluasi Explainability dengan Grad-CAM

Penilaian Expert Simple Assessment terhadap 80 heatmap menunjukkan 57,5% mendapat skor 2 (BAIK), 27,5% skor 1 (CUKUP), dan 15% skor 0 (BURUK). Kelas meningioma memperoleh performa explainability terbaik (rata-rata skor 1,80 dari 2; 85% BAIK), konsisten dengan karakteristik anatomisnya sebagai tumor ekstra-aksial di tepi otak. Sebaliknya, kelas glioma memperoleh performa terburuk (rata-rata skor 0,90 dari 2; 30% BURUK) karena tidak ditemukan pola spasial khas akibat variasi lokasi tumor yang tinggi.



Gambar 3. Expert Simple Assessment per Kelas (kiri) dan Anomali Confidence Score Glioma (kanan)

Temuan yang paling mengkhawatirkan berasal dari analisis Foreground Ratio (FR) pada glioma: FR saat prediksi salah (0,9472) justru lebih tinggi dibandingkan saat prediksi benar (0,9258), yang berarti heatmap tampak wajar secara visual meski prediksinya keliru sehingga tidak memberikan sinyal peringatan. Temuan ini diperkuat oleh anomali confidence score yang terbalik pada kelas glioma, di mana confidence saat prediksi salah (rata-rata 0,900) lebih tinggi dibandingkan saat benar (0,820), mengindikasikan bahwa nilai confidence model tidak terkalibrasi dengan baik dan tidak dapat dijadikan indikator keandalan prediksi untuk kelas ini.

Analisis Misprediksi

Dari 80 sampel Grad-CAM, terdapat 18 kasus misprediksi (22,5%), dengan 11 di antaranya merupakan kasus high-confidence wrong (confidence $\geq 80\%$) dan seluruhnya berasal dari kelas glioma. Analisis visual heatmap pada seluruh kasus misprediksi mengungkap tiga tipe

kegagalan: Tipe 1 “Correct Attention, Wrong Label” (3 kasus), di mana lokalisasi heatmap tepat namun interpretasi kelas keliru; Tipe 2 “High-Confidence Blind Spot” (11 kasus, seluruhnya glioma), kondisi berisiko klinis tertinggi karena heatmap tampak meyakinkan meski prediksi salah; dan Tipe 3 “Complete Spatial Failure” (3 kasus pituitary yang diprediksi meningioma), di mana model kehilangan orientasi spasial sepenuhnya namun anomalnya mudah dikenali secara visual. Kasus paling kritis secara klinis adalah sebuah citra glioma yang diklasifikasikan sebagai no_tumor dengan confidence 99,2%, yang dalam skenario klinis nyata berpotensi menyebabkan pasien dinyatakan bebas tumor padahal mengidap glioma aktif.

Sebagai eksperimen tambahan, pengujian Test-Time Augmentation (TTA) justru menurunkan F1-macro dari 0,8595 menjadi 0,7915 (−6,79%), karena augmentation layer yang dirancang untuk pelatihan tidak sesuai digunakan dalam konteks inference averaging. Temuan ini merupakan eksperimen negatif yang valid, sehingga model base tanpa TTA dikonfirmasi sebagai model final.

Pembahasan

Secara keseluruhan, strategi fine-tuning selektif pada EfficientNetB0 terbukti meningkatkan performa klasifikasi tumor otak hingga melampaui target KPI. Namun, apabila dievaluasi hanya dari nilai akurasi makro, model seolah-olah siap diimplementasikan secara klinis. Analisis Grad-CAM membuka “kotak hitam” tersebut dan mengungkap cacat laten yang kritis pada deteksi glioma, mencakup F1-score rendah, confidence score yang tidak terkalibrasi, serta pola misprediksi berisiko tinggi yang sulit dideteksi secara visual. Temuan ini sejalan dengan urgensi penerapan XAI dalam domain medis sebagaimana ditekankan dalam literatur, bahwa transparansi model tidak boleh diasumsikan hanya dari metrik performa agregat. Kolaborasi antara output sistem CAD dan keahlian radiolog (human-in-the-loop) tetap menjadi prasyarat mutlak sebelum sistem semacam ini dapat diadopsi secara penuh di fasilitas layanan medis.

KESIMPULAN

Penelitian ini berhasil membangun sistem klasifikasi tumor otak berbasis EfficientNetB0 dengan strategi fine-tuning selektif yang mencapai F1-macro 0,8595 pada test set, melampaui target KPI 0,85, serta berhasil mengimplementasikan Grad-CAM secara mandiri untuk menghasilkan dan mengevaluasi 80 heatmap explainability. Evaluasi kualitatif dan kuantitatif menunjukkan bahwa transparansi model bervariasi signifikan antar kelas, dengan kelas glioma teridentifikasi sebagai kelemahan sistemik paling kritis akibat F1-score rendah, anomali confidence score yang terbalik, dan sebelas kasus high-confidence wrong yang seluruhnya berasal dari kelas ini. Temuan ini menegaskan bahwa sistem AI pada kondisi saat ini belum aman digunakan secara mandiri untuk mendeteksi glioma tanpa pengawasan klinis, dan direkomendasikan penerapan augmentasi data spesifik glioma, kalibrasi confidence model (misalnya Temperature Scaling), pengembangan dataset beranotasi lokasi tumor, serta validasi silang dengan metode XAI lain seperti SHAP atau LIME pada penelitian selanjutnya.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Windra Swastika, Ph.D. selaku Pimpinan Laboratorium AiDiTech dan Dosen Pembimbing atas bimbingan dan kesempatan magang riset yang diberikan, kepada Universitas Ma Chung serta Laboratorium AiDiTech atas dukungan fasilitas penelitian, serta kepada seluruh pihak yang telah membantu kelancaran pelaksanaan penelitian ini.

DAFTAR PUSTAKA

- Bhuvaji, S., Kadam, A., Bhumkar, P., Dedge, S., & Kanchan, S. (2020). Brain Tumor Classification (MRI) [Data set]. Kaggle. <https://doi.org/10.34740/KAGGLE/DSV/1183165>
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 248–255. <https://doi.org/10.1109/CVPR.2009.5206848>
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 618–626. <https://doi.org/10.1109/ICCV.2017.74>
- Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. Proceedings of the 36th International Conference on Machine Learning (ICML), 6105–6114.
- Tjoa, E., & Guan, C. (2021). A survey on explainable artificial intelligence (XAI): Toward medical XAI. IEEE Transactions on Neural Networks and Learning Systems, 32(11), 4793–4813. <https://doi.org/10.1109/TNNLS.2020.3027314>
- Zuiderveld, K. (1994). Contrast limited adaptive histogram equalization. In P. S. Heckbert (Ed.), Graphics Gems IV (pp. 474–485). Academic Press.