

Four-Class Brain Tumor Classification Using ConvNeXt with Grad-CAM-Based Explainable

Ni Kadek Jegeg Anastasya Dwipayanti^{1*}, I Gusti Ngurah Lanang Wijayakusuma^{2*}

^{*} Matematika, Universitas Udayana

dwipayanti.2308541038@student.unud.ac.id¹, lanang_wijaya@unud.ac.id²

Article Info

Article history:

Received 2026-07-08

Revised 2026-07-27

Accepted 2026-08-12

Keyword:

CNN,
MRI Classification,
ConvNeXt,
Brain Tumor,
Grad-CAM.

ABSTRACT

Brain tumor is one of the most critical diseases with high mortality rates requiring early and accurate diagnosis. This study proposes a deep learning-based brain tumor classification system using the ConvNeXt-Base architecture to classify four categories: glioma, meningioma, pituitary tumor, and no tumor. To further advance this field, this study addresses two open challenges in prior literature, the need for larger and more diverse datasets to ensure model robustness, and the demand for model transparency (resolving the black-box problem) to facilitate clinical adoption.. To address these, this study combines MRI images from three public sources, Figshare (Cheng et al., 2016), Br35H (Hamada, 2020), and Mendeley (Hira et al., 2025), totaling 11,474 unique images after perceptual hash-based deduplication. The model was trained using progressive unfreezing with AdamW optimizer, CosineAnnealingLR scheduler, and class-weighted cross-entropy loss. Gradient-weighted Class Activation Mapping (Grad-CAM) is integrated as an Explainable AI (XAI) approach to visualize the regions influencing model decisions. The proposed ConvNeXt-Base model achieved an accuracy of 99.25%, macro precision of 99.23%, macro recall of 99.22%, macro F1-score of 99.22%, and macro AUC-ROC of 0.9995 on the test set. Grad-CAM visualizations confirm that the model focuses on clinically relevant tumor regions, particularly the sella turcica for pituitary tumors and distinct mass boundaries for meningioma, thereby improving interpretability and clinical trustworthiness of the AI system.



This is an open access article under the [CC-BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.

I. PENDAHULUAN

Penyakit tumor otak merupakan pertumbuhan sel yang tidak normal dan tidak terkendali di dalam jaringan otak yang dapat berakibat fatal bagi pasien. Tumor otak, baik yang bersifat jinak maupun ganas seperti glioma, meningioma, dan tumor pituitary, menyumbang angka mortalitas yang signifikan secara global, terutama di kalangan anak-anak dan dewasa muda [1]. Diagnosis dini menjadi faktor krusial untuk meningkatkan tingkat kelangsungan hidup pasien melalui intervensi medis yang cepat. Saat ini, pemindaian *Magnetic Resonance Imaging* (MRI) adalah standar emas dalam deteksi tumor otak karena sifatnya yang non-invasif dan kemampuannya menghasilkan resolusi tinggi pada jaringan lunak. Namun, interpretasi manual oleh ahli radiologi masih menghadapi kendala besar, yaitu subjektivitas, variabilitas

antar-pengamat, serta risiko adanya kesalahan diagnosis karena faktor yang kompleks [2].

Beberapa tahun terakhir, teknologi *Deep Learning*, khususnya *Convolutional Neural Networks* (CNN), telah menunjukkan performa luar biasa dalam otomatisasi analisis citra medis. Sejumlah penelitian telah mengeksplorasi berbagai arsitektur untuk mengatasi tantangan ini. Misalnya, penelitian yang mengembangkan kerangka kerja *dual-stage deep learning* menggunakan NASNetMobile dan mencapai akurasi tertinggi (97%) dalam klasifikasi dan lokalisasi tumor otak. Namun, ketergantungan pada dataset tunggal (TCGA-GBM) dari satu sumber membatasi kemampuan generalisasi model pada kondisi klinis nyata [3]. Selanjutnya, penelitian oleh [4] menggunakan arsitektur Convolutional Neural Network (CNN) VGG16 yang telah melalui proses *pretraining*. Model yang dikembangkan mampu mencapai akurasi klasifikasi yang sangat tinggi, yaitu 99,24%. Namun,

penelitian tersebut belum mengimplementasikan aspek interpretabilitas sehingga belum dapat memberikan penjelasan mengenai bagian citra yang menjadi dasar pengambilan keputusan model. Lebih lanjut penelitian oleh [5] menunjukkan bahwa *Explainable AI* yaitu Grad-CAM efektif dalam menjelaskan keputusan model melalui visualisasi area penting pada citra MRI. Akan tetapi, penelitian tersebut hanya mengimplementasikannya pada arsitektur CNN klasik yaitu VGG16 dan Xception, sehingga efektivitas Grad-CAM pada arsitektur CNN modern seperti ConvNext masih menjadi celah penelitian yang perlu dikaji. Dalam aplikasi medis, interpretabilitas merupakan aspek penting karena memungkinkan tenaga medis memahami alasan di balik prediksi model. Grad-CAM menyediakan visualisasi area tumor yang menjadi dasar klasifikasi sehingga meningkatkan transparansi dan mendukung validasi keputusan klinis [6].

Untuk mengatasi celah tersebut, penelitian ini mengusulkan penerapan arsitektur ConvNeXt Base. ConvNeXt Base merupakan evolusi modern CNN yang mengadopsi prinsip desain *Transformer* (seperti *depthwise convolutions*) untuk mencapai performa terdepan dengan stabilitas yang lebih tinggi daripada model *Vision Transformer* konvensional pada dataset citra medis [7]. Berdasarkan penelitian terbaru, ConvNeXt mampu mencapai akurasi 99,13% dan terbukti 53% lebih rendah dalam potensi kesalahan klasifikasi dibandingkan model ViT [8]. Sejalan dengan pentingnya interpretabilitas dalam aplikasi AI di bidang medis, penelitian ini mengintegrasikan metode Explainable Artificial Intelligence (XAI) menggunakan Grad-CAM. Teknik ini menghasilkan visualisasi heatmap yang menunjukkan wilayah citra MRI yang paling memengaruhi keputusan model, sehingga meningkatkan transparansi, mendukung validasi klinis, dan membantu membangun kepercayaan terhadap hasil diagnosis berbasis AI [9].

Penelitian ini diharapkan dapat memberikan kontribusi dalam pengembangan sistem *Computer-Aided Diagnosis* (CAD) yang memiliki performa klasifikasi tinggi sekaligus mampu menyajikan interpretasi visual terhadap proses pengambilan keputusan model. Dengan demikian, sistem yang dikembangkan dapat menjadi alat bantu bagi tenaga medis, khususnya radiolog, dalam mendukung proses analisis citra MRI tumor otak secara lebih akurat, transparan, dan efisien.

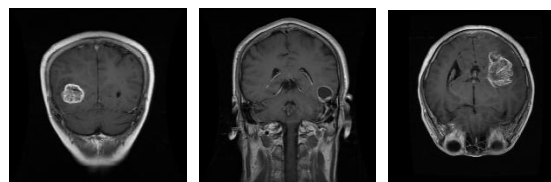
II. METODE

Metodologi dalam penelitian ini dirancang secara sistematis melalui beberapa tahapan utama: pengumpulan data, pra-pemrosesan, augmentasi data, perancangan model, serta evaluasi performa

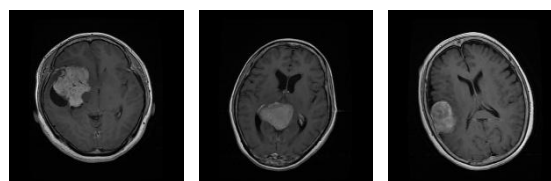
A. Pengumpulan Dataset

Penelitian ini menggunakan dataset citra MRI otak gabungan yang bersumber dari repositori publik yaitu Figshare Brain Tumor Dataset [10] yang berisi 3.064 citra MRI dalam format MATLAB dengan tiga kelas tumor, Br35H Brain Tumor Detection Dataset dari Kaggle yang

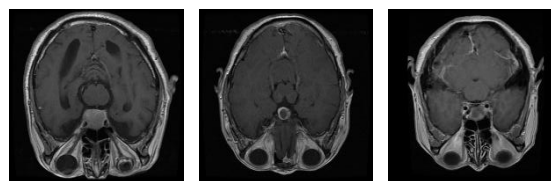
diamil khusus untuk dataset no_tumor sebanyak 2.000 citra [11], dan Brain Tumor MRI Dataset dari Mendeley yang berisi 12.064 citra yang dibagi 4 kelas [12]. Jumlah dataset yang terkumpul dari 3 gabungan dataset tersebut sejumlah 17.128 data. Dataset mencakup empat kategori utama: Glioma, Meningioma, Pituitary, dan Tanpa Tumor (*Normal*).



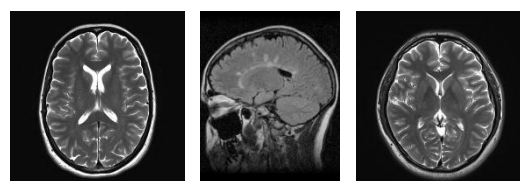
Gambar 1. Citra MRI Glioma



Gambar 2. Citra MRI Meningioma



Gambar 3. Citra MRI Pituitary



Gambar 4. Citra MRI No Tumor

Dari dataset mentah yang dikumpulkan dilakukan deduplikasi untuk menghapus citra yang memiliki duplikasi dan kemiripan identik. Setelah dilakukan deduplikasi diperoleh jumlah dataset unik sebanyak 11.474 dan jumlah dataset yang dihapus sebanyak 5.654 citra. Berikut adalah tabel distribusi dataset unik yang akan digunakan untuk *training*, *validation*, dan *testing*.

TABEL I
DISTRIBUSI DATASET

Kelas	Total Bersih	Train (70%)	Val (15%)	Test (15%)
Glioma	4.314	3.091	647	648
Meningioma	2.554	1.788	383	383
Pituitary	3.102	2.171	465	466
No Tumor	1.504	1.053	226	225
Total	11.474	8.031	1.721	1.722

B. Pra-pemrosesan Citra

Dalam penelitian ini digunakan dataset Figshare yang Dimana format citra dari sumber ini tersimpan dalam format

MATLAB (.mat), sehingga perlu dikonversi dari .mat ke JPEG dengan struktur cdata yang memuat array piksel dan label numerik (1 = meningioma, 2 = glioma, 3 = pituitary). Setiap file dikonversi dengan normalisasi nilai piksel ke rentang 0-255 menggunakan normalisasi min-mx, kemudian disimpan sebagai JPEG dengan kualitas 95%.

Beberapa citra yang dikumpulkan tersimpan dalam mode grayscale atau memiliki channel alpha (RGBA) sehingga perlu distandaisasi format ke RGB untuk memastikan konsistensi input yang dibutuhkan ConvNeXt-Base. Kemudian dilakukan deduplikasi perceptual hashing, pHash dengan hash-size=8 diterapkan pada seluruh 17.128 citra duplikat ditemukan dan dihapus, dan menghasilkan 11.474 citra unik. Prioritas dipertahankan berdasarkan kredibilitas sumber, yaitu yang utama Figshare kemudian Mendeley Data, terakhir Br35H. Seluruh citra yang akan digunakan di resize ke 224x224 piksel, langkah ini dilakukan untuk memastikan input data yang seragam, memperkecil skala, serta dapat mempercepat proses konvergensi model [13]. Kemudian setelah citra dikonversi menjadi tensor dengan rentang nilai [0,1], setiap channel RGB dinormalisasi menggunakan nilai mean (0.485, 0.456, 0.406) dan standar deviasi (0.229, 0.224, 0.225) yang berasal dari dataset ImageNet. Normalisasi ini bertujuan menyelaraskan distribusi data masukan dengan distribusi data yang digunakan pada tahap pre-training model ConvNeXt, sehingga proses transfer learning dapat berlangsung secara optimal.

C. Augmentasi Data

Augmentasi data diterapkan hanya pada set training untuk meningkatkan keberagaman data dan mengurangi Risiko overfitting Transformasi yang digunakan meliputi:

- 1) *Random Horizontal Flip*: citra dibalik secara horizontal secara acak dengan probabilitas 50%. Hal ini diterapkan karena citra MRI otak memiliki karakteristik simetris bilateral yang kuat antara hemisfer kiri dan kanan. Dengan menerapkan rotasi pada rentang sudut tertentu, model dilatih untuk mempertahankan konsistensi diagnostik terhadap pergeseran orientasi spasial, sehingga mampu mengenali fitur tumor meskipun terjadi penyimpangan dari bidang mid-sagital ideal akibat kesalahan pemosisian teknis[14].
- 2) *Random Vertical Flip*: sama halnya dengan random horizontal flip, citra di balik secara vertikal dan secara acak dengan probabilitas 50%, hal ini ditujukan agar model dapat beradaptasi terhadap perubahan orientasi citra serta meningkatkan kemampuan generalisasi terhadap berbagai posisi fitur yang berbeda [8].
- 3) *Random Rotation*: Teknik rotasi dalam rentang -15° hingga $+15^\circ$ diterapkan untuk meningkatkan kemampuan model dalam mengenali objek meskipun terjadi variasi orientasi citra [15]
- 4) *Brightness dan Contrast Adjustment*: karena dataset MRI yang digunakan grayscale, maka augmentasi untuk

mengubah tingkat kecerahan dan kontras dilakukan hingga $\pm 20\%$ menggunakan transformasi ColorJitter. Tujuan augmentasi ini adalah untuk meningkatkan ketahanan model terhadap variasi intensitas dan kontras citra MRI yang muncul sehingga kemampuan generalisasi model terhadap data dari berbagai sumber menjadi lebih baik.

D. Perancangan Model Arsitektur ConvNeXt

Penelitian ini mengimplementasikan arsitektur ConvNeXt, yang dipilih karena kemampuannya menggabungkan efisiensi konvolusional CNN dengan prinsip desain modern dari *Vision Transformer*. ConvNeXt secara khusus efektif untuk dataset medis karena memiliki *inductive bias* spasial yang kuat dan efisiensi parameter yang lebih baik dibandingkan *Vision Transformer* (ViT) konvensional.

Model yang digunakan adalah ConvNeXt-Base yang telah melalui proses pretraining menggunakan dataset ImageNet-1K. Pada tahap transfer learning, lapisan klasifikasi bawaan ConvNeXt diganti dengan lapisan Fully Connected (Linear Layer) baru yang memiliki jumlah neuron keluaran sesuai dengan jumlah kelas pada dataset penelitian yaitu glioma, meningioma, pituitary, dan no tumor.

Proses pelatihan model dilakukan menggunakan strategi dua tahap (two-stage fine tuning). Tahap pertama adalah head training, yaitu membekukan seluruh parameter backbone ConvNeXt sehingga hanya lapisan klasifikasi yang diperbarui selama proses pelatihan. Tahap ini bertujuan untuk menyesuaikan classifier baru terhadap karakteristik dataset MRI tanpa mengubah representasi fitur dasar yang telah dipelajari selama proses pretraining.

Setelah classifier memiliki kemampuan klasifikasi awal yang baik, proses dilanjutkan ke tahap kedua yaitu full fine-tuning. Pada tahap ini seluruh parameter model dibuka kembali sehingga backbone ConvNeXt dan classifier diperbarui secara bersamaan. Strategi ini memungkinkan model menyesuaikan representasi fitur secara lebih spesifik terhadap karakteristik citra MRI tumor otak sehingga menghasilkan performa klasifikasi yang lebih optimal.

Optimasi parameter dilakukan menggunakan algoritma AdamW, yang merupakan pengembangan dari Adam dengan mekanisme weight decay terpisah sehingga mampu meningkatkan kemampuan generalisasi model. Fungsi objektif yang digunakan adalah *Cross Entropy Loss* dengan penerapan label smoothing sebesar 0,1 untuk mengurangi kecenderungan model menghasilkan prediksi yang terlalu yakin (*overconfident*).

Selain itu, laju pembelajaran (learning rate) diatur menggunakan Cosine Annealing Learning Rate Scheduler. Scheduler ini secara bertahap menurunkan nilai learning rate mengikuti fungsi kosinus selama proses pelatihan sehingga pembaruan parameter menjadi lebih stabil ketika model mendekati titik konvergensi.

Untuk menghindari overfitting, penelitian ini menerapkan mekanisme Early Stopping berdasarkan nilai akurasi validasi. Apabila akurasi validasi tidak mengalami peningkatan selama sejumlah epoch tertentu (patience), maka proses pelatihan

dihentikan dan model dengan performa validasi terbaik disimpan sebagai model akhir.

Evaluasi model dilakukan menggunakan metode 5-Fold Cross Validation, yaitu dataset dibagi menjadi lima subset dengan setiap subset secara bergantian digunakan sebagai data validasi, sedangkan empat subset lainnya digunakan sebagai data pelatihan. Pendekatan ini menghasilkan evaluasi yang lebih stabil dan mengurangi ketergantungan performa model terhadap pembagian data tertentu [16].

TABEL II
HYPERPARAMETER YANG DIGUNAKAN

Hyperparameter	Detail
Optimizer	AdamW
Learning Rate Tahap 1	1×10^{-3}
Learning Rate Tahap 2	1×10^{-4}
Wight Decay	0,05
Batch Size	32
Early Stopping Patience	10 Epoch
Head Training	5 Epoch
Fine Tuning	Maksimum 50 Epoch

E. Implementasi Explainable AI (XAI)

Penelitian ini menerapkan pendekatan Explainable Artificial Intelligence (XAI) untuk meningkatkan interpretabilitas hasil klasifikasi. Metode XAI yang digunakan dalam penelitian ini adalah Gradient-wighted Class Activation Mapping (Grad-CAM). Implementasi Grad-CAM dilakukan setelah proses pelatihan selesai menggunakan model dengan performa terbaik berdasarkan akurasi validasi.

Citra MRI yang akan dianalisis terlebih dahulu melalui tahap prapemrosesan yang sama seperti pada proses pelatihan, selanjutnya dimasukkan ke dalam model ConvNeXt-Base untuk memperoleh probabilitas prediksi setiap kelas menggunakan *Softmax*. Kelas dengan probabilitas tertinggi dipilih sebagai kelas target yang akan dijelaskan menggunakan Grad-CAM.

Pada tahap berikutnya, Grad-CAM menghitung gradien dari skor kelas target terhadap feature map pada lapisan konvolusi terakhir (target layer) dari backbone ConvNeXt. Gradien tersebut digunakan untuk menghitung bobot setiap feature map sehingga diperoleh peta aktivasi yang merepresentasikan kontribusi masing-masing fitur terhadap prediksi model. Peta aktivasi kemudian dinormalisasi dan diperbesar (upsampling) agar memiliki ukuran yang sama dengan citra masukan.

Heatmap hasil Grad-CAM selanjutnya ditumpangkan (overlay) pada citra MRI asli sehingga menghasilkan visualisasi yang menunjukkan area yang menjadi fokus perhatian model selama proses klasifikasi. Visualisasi ini digunakan untuk mengevaluasi apakah model memanfaatkan karakteristik anatomis yang sesuai dengan lokasi tumor atau justru dipengaruhi oleh fitur yang tidak relevan.

Pada penelitian ini, analisis Grad-CAM dilakukan terhadap dua kelompok citra, yaitu:

- 1) Sampel yang berhasil diklasifikasikan dengan benar.

- 2) Sampel yang mengalami salah klasifikasi.

Hasil visualisasi tersebut kemudian digunakan sebagai dasar interpretasi terhadap perilaku model serta unut analisis performa model.

Bobot untuk feature map ke-k, yaitu:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \quad (1)$$

Dengan :

- α_k^c = bobot feture map ke-k untuk kelas c
- y^c = skor sebelum Softmax untuk kelas target
- A_{ij}^k = aktivasi pada feature map ke-k
- Z = jumlah piksel pada feature map

Selanjutnya, peta aktivasi Grad-CAM dihitung menggunakan:

$$L_{Grad-CAM}^c = ReLU \sum_i \alpha_k^c A^k \quad (2)$$

Fungsi ReLU digunakan untuk mempertahankan hanya kontribusi positif terhadap kelas target sehingga heatmap yang dihasilkan menunjukkan wilayah citra yang benar-benar mendukung Keputusan model [17].

F. Metrik Evaluasi

Untuk mengukur performa model secara objektif, dilakukan evaluasi model dengan berbagai metrik yang relevan. Evaluasi dilakukan untuk menilai kemampuan model dalam generalisasi dan memastikan kestabilan performa secara menyeluruh pada berbagai kondisi dan variasi dataset. Evaluasi disajikan dalam bentuk tabel dan grafik agar lebih mudah dianalisis dan dibandingkan. Dalam penelitian ini metrik evaluasi yang digunakan meliputi Accuracy, Presicion, Recall, F1-Score, dan AUC-ROC. Kelima metrik tersebut dihitung berdasarkan nilai pada confusion matrix, yang terdiri atas *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN).

Keterangan confusion matrix Adalah sebagai berikut:

- *True Positive* (TP), Adalah jumlah sampel positif yang berhasil diklasifikasikan dengan benar oleh model.
- *True Negative* (TN), Adalah jumlah positif yang berhasil diklasifikasikan dengan benar oleh model.
- *False Positive* (FP), adalah jumlah sampel negatif yang salah diprediksi sebagai positif.
- *False Negative* (FN): adalah jumlah sampel positif yang salah diprediksi sebagai negatif.

1) *Accuracy*: Merupakan metrik yang digunakan merupakan metrik yang digunakan untuk mengukur proporsi keseluruhan orediksi yang benar terhadap seluruh data yang diuji. Metrik ini memberikan gambaran umum mengenai performa model dalam melakukan klasifikasi. Nilai accuracy yang tinggi menunjukkan bahwa sebagian besar sampel berhasil diklasifikasikan dengan benar. Namun, pada dataset yang memiliki distribusi kelas tidak seimbang, accuracy saja

belum cukup untuk menggambarkan kemampuan model secara menyeluruh.

Rumus accuracy dinyatakan sebagai berikut.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

2) *Precision*: Mengukur Tingkat ketepatan model dalam memberikan prediksi positif. Metrik ini menunjukkan proporsi prediksi positif yang benar dibandingkan dengan seluruh sampel yang diprediksi sebagai positif. Nilai precision yang tinggi menunjukkan bahwa model memiliki tingkat kesalahan False Positive yang rendah.

Rumus precision adalah sebagai berikut.

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

3) *Recall*: Mengukur kemampuan model dalam mendeteksi seluruh sampel yang benar-benar termasuk ke dalam kelas positif. Metrik ini menunjukkan seberapa banyak data positif yang berhasil dikenali model. Pada aplikasi diagnosis medis, recall menjadi salah satu metrik yang penting karena berkaitan dengan kemampuan model dalam meminimalkan jumlah False negative, yaitu kasus penyakit yang tidak berhasil terdeteksi.

Rumus recall dinyatakan sebagai berikut.

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

4) *F1-Score*: Merupakan rata-rata harmonik antara nilai precision dan recall. Metrik ini digunakan untuk memberikan keseimbangan antara kemampuan model dalam menghasilkan prediksi positif yang tepat serta kemampuannya mendeteksi seluruh sampel positif. Nilai F1-Score menjadi indikator yang penting ketika kedua metrik tersebut sama-sama perlu diperhatikan.

Rumus F1-Score adalah sebagai berikut

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (6)$$

5) *Area Under the Receiver Operating Characteristic Curve (AUC-ROC)*: Merupakan metrik yang digunakan untuk mengevaluasi kemampuan model dalam membedakan antara kelas positif dan kelas negatif pada berbagai nilai ambang (threshold). Kurva ROC dibentuk dari hubungan antara True Positive Rate (TPR) dan False Positive Rate (FPR). Semakin mendekati nilai 1, semakin baik kemampuan model dalam membedakan kedua kelas, sedangkan nilai 0,5 menunjukkan bahwa performa model setara dengan prediksi secara acak.

Kurva ROC dibentuk menggunakan persamaan berikut:

$$TPR = \frac{TP}{TP+FN} \quad (7)$$

$$FPR = \frac{FP}{FP+TN} \quad (8)$$

Selanjutnya, nilai AUC diperoleh dari luas daerah di bawah kurva ROC. Nilai AUC berada pada rentang 0 hingga 1, di mana nilai yang lebih tinggi menunjukkan kemampuan diskriminasi model yang semakin baik.

III. HASIL DAN PEMBAHASAN

A. Performa Klasifikasi Keseluruhan

Tabel II menyajikan perbandingan performa keseluruhan antara model yang diusulkan (Convnext-Base) dan model pembanding (Resnet-50) pada test set yang terdiri dari 1.722 citra MRI yang tidak digunakan pada model selama proses pelatihan.

TABEL III
PERBANDINGAN PERFORMA MODEL

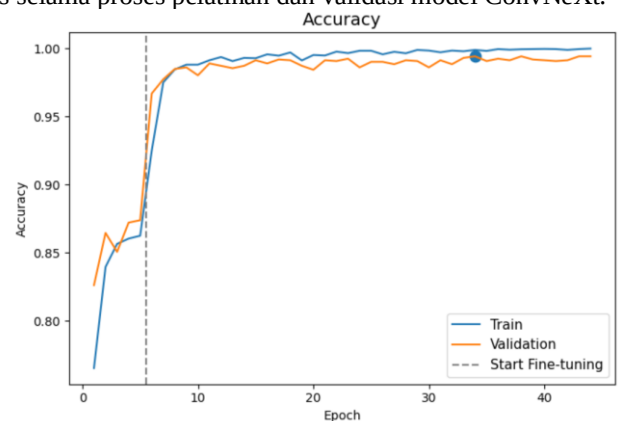
Model	Acc	Preci sion	Recall	F1- Score	AUC- ROC
Convnext	99,25%	99,23%	99,22%	99,22%	0,9995
Resnet-50	99,30%	99,22%	99,25%	99,23%	0,9991

Hasil pada tabel menunjukkan bahwa Convnext-Base mencapai akurasi 99,25% dengan macro F1-score 98,22% dan macro AUC-ROC 0,9995 pada test set. Akurasi yang tinggi menunjukkan bahwa model mampu mengklasifikasikan hampir seluruh citra MRI dengan benar.

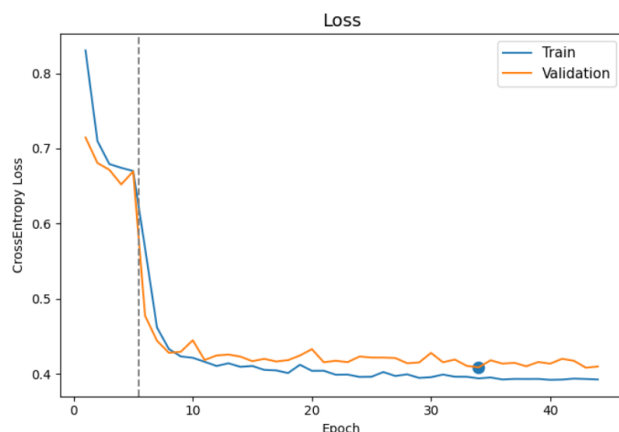
Secara kuantitatif, Resnet-50 menunjukkan angka yang sedikit lebih tinggi pada seluruh matrik evaluasi, dengan selisih accuracy sebesar 0,05 % dan F1-Score sebesar 0,01%, perbedaan ini tidak substansial secara praktis pada performa *near-ceiling* di atas 99%.

Dalam konteks ini pemilihan ConvNeXt-Base sebagai model utama penelitian ini didasarkan pada pertimbangan bahwa desain large kernel depthwise convolution 7x7 dibandingkan kernel 3x3 pada ResNet50, sehingga model ConvNeXt-Base secara arsitektural memberikan kapasitas pemrosesan spasial yang lebih kuat, dan lebih unggul dalam segmentasi dan deteksi objek [18].

Berikut akan ditunjukkan nilai perkembangan akurasi dan loss selama proses pelatihan dan validasi model ConvNeXt.



Gambar 5. Kurva Accuracy ConvNext



Gambar 6. Kurva Loss ConvNext

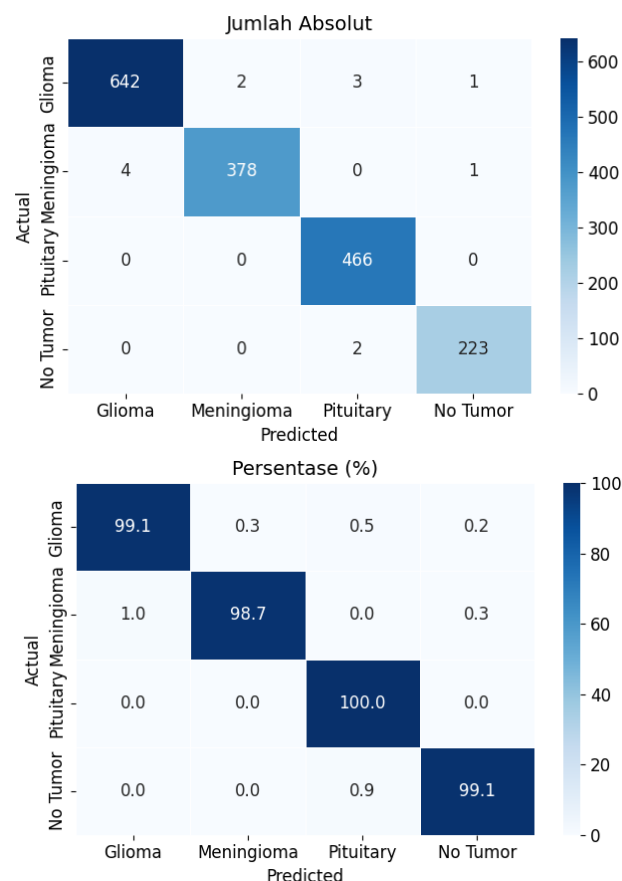
Dari kedua hasil tersebut diperoleh bahwa setelah memasuki tahap fine-tuning, terjadi peningkatan accuracy maupun penurunan loss yang signifikan sehingga tidak ditemukan indikasi overfitting pada model. Sekitar epoch ke-10, hasil kurva menunjukkan nilai accuracy dan loss yang relatif stabil dengan fluktuasi yang sangat kecil. Nilai yang ditunjukkan dalam kurva sesuai dengan nilai accuracy yang mencapai 99,30%. Secara keseluruhan, kurva accuracy dan loss menunjukkan bahwa strategi transfer learning dua tahap, yaitu feature extraction dan dilanjutkan dengan fine-tuning, efektif dalam meningkatkan kinerja ConvNeXt-Base pada klasifikasi citra MRI otak.

B. Analisis Performa Per Kelas

Tabel III menyajikan metrik evaluasi ConvNeXt-Base secara rinci per kelas, sementara Gambar 1 menampilkan confusion matrix dalam bentuk nilai absolut dan persentase.

TABEL IV
PERBANDINGAN PERFORMA MODEL

Model	Precision	Recall	F1-Score	AUC-ROC	n/salah
Glioma	99,38%	99,07%	99,23%	0,9999	648/6
Meningioma	99,47%	98,69%	99,08%	0,9999	383/5
Pituitary	98,94%	100,00%	99,47%	1,000	466/0
No Tumor	99,11%	99,11%	99,11%	0,9984	225/2
Macro AVG	99,23%	99,22%	99,22%	0,9995	1.722/13



Gambar 7. Confusion Matrix ConvNeXt-Base

Kelas Pituitary menunjukkan performa terbaik dengan F1-score tertinggi sebesar 99,47% dan AUC-ROC sebesar 1,000 serta tidak terdapat kesalahan uji dari 466 citra. Hal ini dapat dijelaskan oleh karakteristik morfologi Pituitary yang secara anatomis terletak pada area sella turcica, sehingga memberikan konteks spasial yang paling jelas dan paling mudah dipelajari oleh model.

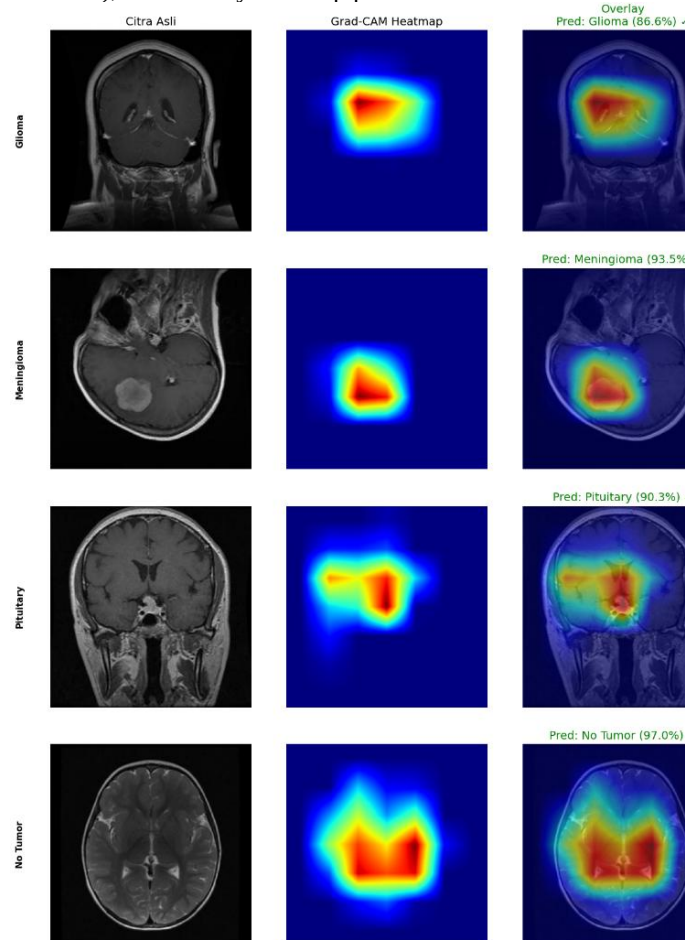
Kelas Meningioma menunjukkan recall terendah sebesar 98,69% dan total 5 kesalahan prediksi. Analisis detail confusion matrix mengungkapkan bahwa dari 5 kesalahan tersebut, 4 citra meningioma salah prediksi sebagai glioma dan 1 salah prediksi sebagai no tumor. Tingginya variabilitas morfologi meningioma yang dapat muncul dengan berbagai bentuk, ukuran, dan lokasi dapat menjadi faktor utama kesulitan klasifikasi ini. Kemiripan visual antara meningioma dan glioma pada irisan MRI tertentu, terutama saat meningioma menginvasi parenkim otak, menjelaskan kebingungan antara kedua kelas ini.

Kelas No Tumor menghasilkan 2 kesalahan prediksi, sebagai pituitary. Secara klinis, kesalahan ini dapat terjadi karena model mendeteksi struktur anatomis normal yang terdapat di area sellar dan suprasellar, seperti kelenjar hipofisis normal, tangkai hipofisis, atau cisterna suprasellar sebagai area patologis. Ketiga kasus ini merupakan limitasi yang paling signifikan secara klinis karena dapat menyebabkan prosedur diagnostic lanjutan yang tidak

diperlukan pada pasien sehat, dan perlu menjadi fokus perbaikan pada penelitian selanjutnya.

C. Visualisasi Interpretabilitas Model dengan Grad-Cam

Gambar 2 menampilkan visualisasi Grad-Cam untuk satu citra representatif dari setiap kelas tumor yang berhasil diprediksi dengan benar oleh ConvNext-Base. Setiap baris terdiri dari tiga panel, yaitu citra MRI asli, heatmap Grad-Cam dalam skala jet (merah = aktivasi tertinggi, biru = terendah), dan overlay heatmap pada citra asli.



Gambar 7. Visualisasi Grad-Cam seluruh kelas oleh ConvNext-Base

Pada visualisasi Glioma heatmap Grad-CAM menunjukkan aktivitas yang tersebar luas di area sentral dan parasentral otak, mencerminkan sifat infiltratif glioma yang tidak memiliki batas tegas. Model berhasil memprediksi kelas ini dengan confidence 86,6%.

Heatmap pada visualisasi meningioma menunjukkan aktivasi yang terfokus pada massa tumor berbentuk bulat dengan batas yang jelas di permukaan otak (konvektivitas serebri). Adanya perbedaan densitas yang jelas antara massa tumor dan jaringan otak normal yang terlihat pada citra asli berhasil diidentifikasi dengan confidence prediksi 91,4%.

Kemudian pada visualisasi heatmap pituitary menunjukkan bahwa aktivasi grad-cam sangat terlokalisasi pada area sella turcica di garis tengah dasar tengkorak, yang merupakan

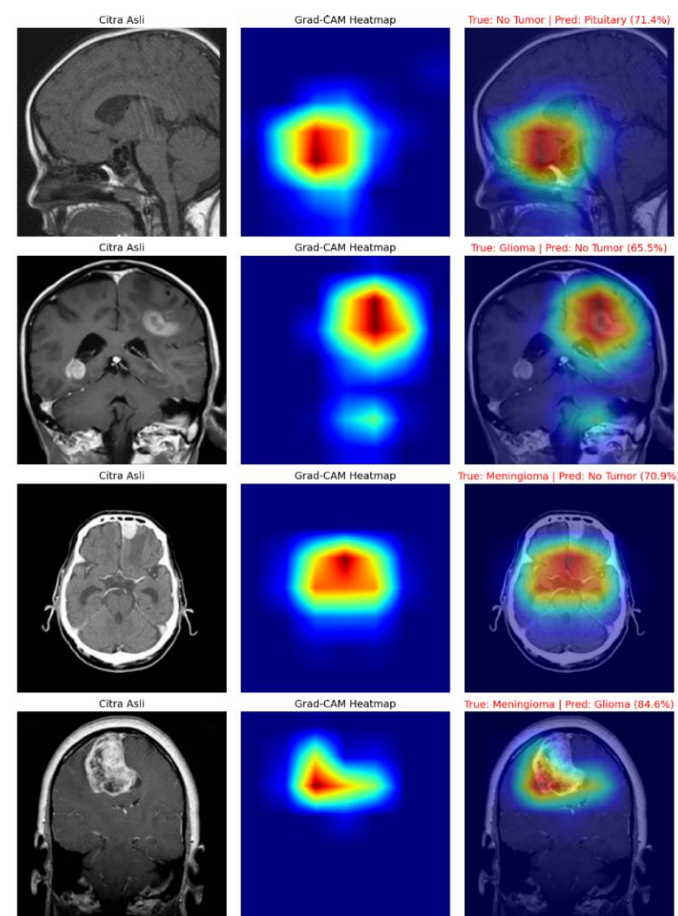
lokasi anatomis khas tumor pituitary. Kelas ini mencapai AUC-ROC tertinggi dengan confidence prediksi 91,4%.

Terakhir untuk kelas no tumor, heatmap pada kelas ini menunjukkan pola aktivasi tersebar difus di seluruh parenkim otak tanpa adanya fokus spesifik pada satu area anomali. Model dengan confidence 96,9% mengidentifikasi distribusi jaringan otak yang homogen dan simetris sebagai karakteristik otak normal, tanpa mendeteksi adanya massa atau area densitas abnormal yang terlokalisasi.

Secara keseluruhan, visualisasi Grad-CAM membuktikan bahwa ConvNeXt-Base tidak hanya mencapai akurasi tinggi secara statistik, tetapi juga belajar merepresentasikan fitur diagnostik yang bermakna secara klinis untuk setiap kelas tumor. Temuan ini menjawab keraguan terhadap prediksi model dengan memberikan bukti kualitatif bahwa model dapat dipercaya untuk dipertimbangkan sebagai sistem pendukung keputusan klinis

D. Analisis Kasus Kegagalan Prediksi (Failure Case Analysis)

Untuk memahami penyebab kegagalan model akan dilakukan analisis empat kasus kesalahan prediksi yang ditemukan pada dataset menggunakan visualisasi Grad-Cam. Gambar 3 akan menampilkan empat kasus tersebut beserta heatmap Grad-Cam yang dihasilkan.



Gambar 8. Visualisasi No Tumor oleh ConvNext-Base

Pada gambar 8 kelas no tumor yang diprediksi sebagai pituitary dengan confidence 71,4%, menunjukkan heatmap Grad-CAM dengan aktivasi model terfokus pada area sellar dan suprasellar, meskipun berasal dari pasien tanpa tumor. Hal ini mengindikasikan bahwa model mendekteksi struktur anatomis normal di area sella turcica sebagai area patologis. Kesalahan jenis ini merupakan false krisis yang kritis secara klinis karena berpotensi menyebabkan kecemasan pada pasien dan prosedur diagnostik lanjutan yang tidak diperlukan. Temuan ini menyarankan perlunya data pelatihan yang lebih beragam untuk variasi normal anatomi sellar.

Untuk kasus Glioma yang diprediksi sebagai no tumor dengan confidence 65,5% menunjukkan bahwa model dalam kondisi ragu (uncertain), bukan salah dengan keyakinan tinggi. Heatmap Grad-CAM menunjukkan pola aktivasi tersebar tidak beraturan dan tidak terlokasi pada area tumor. Confidence yang rendah ini merupakan perilaku yang baik dari perspektif keselamatan klinis, dalam hal ini model yang ragu tidak memberikan prediksi palsu dengan keyakinan tinggi.

Selanjutnya adalah kasus meningioma yang diprediksi sebagai no tumor dengan confidence 70,9%. Heatmap menunjukkan aktivasi yang tersebar luas tanpa fokus spesifik pada lokasi massa tumor yang ada. Inspeksi citra asli menunjukkan kemungkinan bahwa tumor dalam citra ini memiliki ukuran yang sangat kecil atau kontras yang lemah terhadap jaringan sekitar, sehingga menyerupai jaringan otak normal secara visual. Kasus ini menegaskan pentingnya *threshold* kalibrasi confidence dan sistem flagging untuk kasus kasus di mana model perlu di-review secara manual.

Kemudian kasus meningioma yang diprediksi sebagai glioma dengan confidence 84,6%, heatmap Grad-CAM menunjukkan aktivasi yang tepat pada lokasi massa tumor yang terlihat jelas di citra asli artinya model secara akurat mendeteksi keberadaan dan lokasi tumor, tetapi salah dalam menentukan jenisnya. Ini membuktikan bahwa model tidak membuat kesalahan karena pembelajaran pola yang salah (misalignment) atau berfokus pada artefak. Kesalahan terjadi murni karena ambiguitas visual, yaitu karakteristik morfologi tumor pada citra spesifik ini memang memiliki penampakan yang menyerupai glioma, Ini merupakan bukti yang memperkuat validitas model.

Secara keseluruhan, analisis kegagalan prediksi dengan Grad-Cam mengungkapkan tiga pola kesalahan yang berbeda yaitu, (1) false positive akibat ambiguitas anatomi normal, (2) kesalahan pada citra berkualitas rendah dengan confidence rendah yang menunjukkan perilaku model yang sesuai, (3) kesalahan diferensiasi visual antar tumor yang secara morfologi mirip. Ketiga pola ini konsisten dengan keterbatasan inheren klasifikasi MRI yang bahkan dihadapi oleh radiolog manusia, dan memberikan perpaduan konkret untuk pengembangan sistem selanjutnya, yaitu termasuk penambahan data dengan variasi kualitas akuisisi yang lebih beragam dan integrasi informasi klinis tambahan sebagai fitur pendukung.

IV. KESIMPULAN

Penelitian ini berhasil mengembangkan system klasifikasi tumor otak berbasis citra MRI menggunakan arsitektur ConvNeXt-Base yang diintegrasikan dengan teknik Explainable AI berbasis Grad-CAM. Model ConvNeXt yang diusulkan mampu mencapai accuracy 99,25%, macro F1-score 99,22%, dan macro AUC-ROC sebesar 0,9995 pada test set 1.722 citra. Model pembanding ResNet-50 mencapai performa sedikit lebih tinggi (accuracy 99,30% dan macro F1-score 99,23%). Selisih akurasi yang sangat kecil sehingga perbedaan ini tidak substansial secara praktis pada performa model yang near sealing di atas 99%.

Pemilihan model ConvNeXt yang didasarkan pada kualitas interpretabilitas visual Grad-CAM ConvNeXt-Base terbukti secara anatomis dapat melokalisasi region tumor dengan presisi. Analisis kasus kesalahan prediksi mengkonfirmasi bahwa model tidak belajar pola yang salah, melainkan kesulitan diferensiasi visual abstrak jenis tumor yang memang memiliki penampakan serupa.

DAFTAR PUSTAKA

- [1] V. Essianda, A. D. Indrasari, P. Widyastuti, T. Syahla, and R. Rohadi, "Brain Tumor : Molecular Biology, Pathophysiology, and Clinical Symptoms," *J. Biol. Trop.*, vol. 23, no. 4, pp. 260–269, 2023, doi: 10.29303/jbt.v23i4.5585.
- [2] R. N. Asif *et al.*, "Brain tumor detection empowered with ensemble deep learning approaches from MRI scan images," *Sci. Rep.*, vol. 15, no. 1, pp. 1–22, 2025, doi: 10.1038/s41598-025-99576-7.
- [3] D. Rastogi *et al.*, "Dual-stage deep learning framework for brain tumor classification and localization using multimodal MRI scans," *Intell. Med.*, vol. 13, no. February, 2026, doi: 10.1016/j.ibmed.2026.100361.
- [4] Y. Wong, E. L. M. Su, C. F. Yeong, W. Holderbaum, and C. Yang, "Brain tumor classification using MRI images and deep learning techniques," *PLoS One*, vol. 20, no. 5 May, pp. 1–19, 2025, doi: 10.1371/journal.pone.0322624.
- [5] S. A. Shreya, M. A. I. Siddique, and S. Tasnim, "DB-FGA-Net: Dual Backbone Frequency Gated Attention Network for Multi-Class Brain Tumor Classification with Grad-CAM Interpretability," pp. 1–25, 2025, [Online]. Available: <https://arxiv.org/pdf/2510.20299>
- [6] K. K. A. N. D. K. Jyoti, "Recent Advancements in Grad-CAM and variants: Enhancing Brain Tumor Detection, Segmentation, and Classification," *Res. Sq. (Research Square)*, pp. 1–18, 2024, [Online]. Available: <https://doi.org/10.21203/rs.3.rs-5485128/v1>
- [7] O. P. Mmileng, A. Whata, M. Olusanya, and S. Mhlongo, "Application of ConvNeXt with transfer learning and data augmentation for malaria parasite detection in resource-limited settings using microscopic images," *PLoS One*, vol. 20, no. 6 June, pp. 1–39, 2025, doi: 10.1371/journal.pone.0313734.
- [8] G. Ekayanda and M. Rahardi, "Analysis of Deep Learning Algorithms Using ConvNeXt and Vision Transformer for Brain Tumor Disease," *J. Appl. Informatics Comput.*, vol. 9, no. 6, pp. 3295–3304, 2025, doi: 10.30871/jaic.v9i6.11438.
- [9] K. Pant, P. K. Dutta Pramanik, and Z. Zhao, "A Robust ConvNeXt-Based Framework for Efficient, Generalizable, and Explainable Brain Tumor Classification on MRI," *Bioengineering*, vol. 13, no. 2, 2026, doi: 10.3390/bioengineering13020157.
- [10] J. Cheng, "Brain Tumor Dataset," 2024, *Figshare*. doi: <https://doi.org/10.6084/m9.figshare.1512427>.
- [11] A. Hamada, "Br35H :: Brain Tumor Detection," 2020, *Kaggle*.
- [12] M. I. K. Hira, M. S. Hossain, and M. M. A. Bithee, "Brain Tumor MRI Dataset (Glioma, Meningioma, Pituitary, No Tumor)," 2025, *Mendeley Data*. doi: 10.17632/zwr4ntf94j.1.

- [13] D. L. S. Dimara, S. W. Putri, R. Amelia, Z. I. Arishandy, and A. M. Rizki, "Penerapan Convolutional Neural Network (CNN) dalam Klasifikasi Citra MRI untuk Deteksi Tumor Otak Manusia," *KERNEL J. Ris. Inov. Bid. Inform. dan Pendidik. Inform.*, vol. 4, no. 2, pp. 70–77, 2023, doi: 10.31284/j.kernel.2023.v4i2.6960.
- [14] Y. Liu, R. T. Collins, W. E. Rothfus, and M. D., "Automatic Extraction of the Central Symmetry (Mid-Sagittal) Plane from Neuroradiology Images," *Carnegie Mellon Univ. Robot. Inst. Tech. Report, C.*, pp. 1–29, 1996.
- [15] E. Gocerli, *Medical image data augmentation: techniques, comparisons and interpretations*, vol. 56, no. 11. Springer Netherlands, 2023. doi: 10.1007/s10462-023-10453-z.
- [16] B. S. Abd El-Wahab, M. E. Nasr, S. Khamis, and A. S. Ashour, "BTC-fCNN: Fast Convolution Neural Network for Multi-class Brain Tumor Classification," *Heal. Inf. Sci. Syst.*, vol. 11, no. 1, pp. 1–22, 2023, doi: 10.1007/s13755-022-00203-w.
- [17] 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," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, 2020, doi: 10.1007/s11263-019-01228-7.
- [18] Z. Liu, H. Mao, C. Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2022-June, pp. 11966–11976, 2022, doi: 10.1109/CVPR52688.2022.01167.