

© Jurnal Nasional Teknik Elektro dan Teknologi Informasi  
Karya ini berada di bawah Lisensi Creative Commons Atribusi-BerbagiSerupa 4.0 Internasional  
Terjemahan artikel 10.22146/jnteti.v14i4.19822

# Deteksi Pneumonia Menggunakan *Explainable AI*: Model Hibrid CNN–ViT dan Grad-CAM

Atika Hendryani<sup>1</sup>, Vita Nurdinawati<sup>1</sup>, Agus Komarudin<sup>1</sup>

<sup>1</sup> Jurusan Teknik Elektromedik, Poltekkes Kemenkes Jakarta II, Jakarta 12120, Indonesia

[Diserahkan: 28 Mei 2025, Direvisi: 22 Agustus 2025, Diterima: 24 September 2025]

Penulis Korespondensi: Atika Hendryani (email: atika@poltekkesjkt2.ac.id)

**INTISARI** — Deteksi pneumonia melalui citra medis merupakan tantangan signifikan, terutama di wilayah dengan keterbatasan akses terhadap tenaga kesehatan. Penelitian ini menyajikan model *explainable artificial intelligence* (XAI) yang mengintegrasikan *convolutional neural network* (CNN) dan *vision transformer* (ViT) untuk meningkatkan akurasi diagnosis pneumonia berbasis citra rontgen dada (*chest X-ray*). Tujuan penelitian ini adalah meningkatkan ketepatan diagnosis dengan menyediakan penjelasan melalui visualisasi *gradient-weighted class activation mapping* (Grad-CAM). Metodologi penelitian meliputi prapemrosesan citra, ekstraksi fitur lokal menggunakan CNN, serta pemodelan hubungan spasial global melalui ViT. Model dilatih menggunakan *dataset* citra rontgen dada yang telah melalui prapemrosesan dan dievaluasi menggunakan metrik kinerja standar seperti akurasi, presisi, *recall*, dan skor F1. Model hibrid CNN–ViT yang diusulkan diuji menggunakan *dataset* rontgen dada untuk deteksi pneumonia. Hasil eksperimen menunjukkan bahwa model mencapai akurasi sebesar 96,5%, presisi 96%, *recall* 96%, dan skor F1 94%. Hasil ini mengindikasikan bahwa integrasi CNN dan ViT secara efektif meningkatkan kinerja klasifikasi dan menyediakan alat analisis citra medis yang andal. Selain itu, visualisasi Grad-CAM menyoroti area-area penting pada citra yang memengaruhi prediksi model, sehingga meningkatkan interpretabilitas. Dibandingkan dengan model konvensional, pendekatan ini menawarkan transparansi yang lebih baik dalam sistem diagnosis berbasis kecerdasan buatan. Dengan demikian, model yang diusulkan merepresentasikan alat diagnosis yang menjanjikan dan dapat diandalkan, terutama bermanfaat di wilayah terpencil dengan infrastruktur medis terbatas. Penelitian ini juga membuka peluang pengembangan sistem diagnosis yang transparan dan berbasis XAI di masa mendatang.

**KATA KUNCI** — *Explainable AI*, CNN-ViT, Grad-CAM, Pneumonia, Pemrosesan Citra, Citra Medis.

## I. PENDAHULUAN

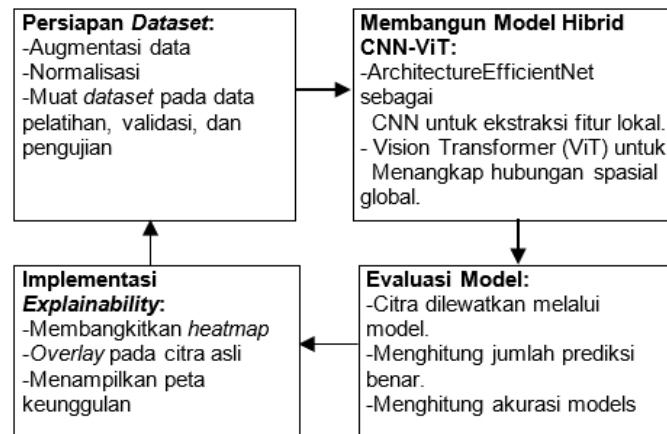
Pneumonia merupakan infeksi pada paru-paru yang menyebabkan peradangan pada alveolus, yaitu kantung udara kecil tempat terjadinya pertukaran gas. Alveolus tersebut dapat terisi oleh cairan atau nanah yang mengganggu proses pertukaran oksigen, sehingga menimbulkan gejala seperti batuk, demam, dan sesak napas. Kondisi ini umumnya disebabkan oleh patogen bakteri seperti *Streptococcus pneumoniae*, virus seperti influenza dan SARS-CoV-2, serta jamur pada individu dengan sistem kekebalan tubuh yang lemah. Dalam praktik klinis, pneumonia umumnya didiagnosis menggunakan teknik pencitraan medis, khususnya rontgen dada (*chest X-ray*), yang membantu dokter mengidentifikasi infiltrasi atau area buram pada paru-paru yang menandakan adanya infeksi.

Metode konvensional dalam analisis citra rontgen dilakukan secara manual oleh radiolog dengan mengenali pola-pola khas pneumonia, seperti opasitas unilateral maupun bilateral pada paru. Perkembangan terkini dalam bidang kecerdasan buatan (*artificial intelligence*, AI) dan teknologi *deep learning* telah memungkinkan identifikasi otomatis terhadap pneumonia menggunakan model berbasis *convolutional neural network* (CNN). Model-model tersebut dilatih dengan menggunakan *dataset* citra rontgen dada berskala besar seperti ChestX-ray dan *dataset* milik Radiological Society of North America (RSNA), untuk mengenali pola pneumonia dengan tingkat akurasi tinggi, sehingga dapat membantu dokter dalam proses diagnosis yang lebih cepat dan akurat.

Diagnosis pneumonia melalui analisis citra toraks telah menjadi pendekatan yang banyak digunakan dalam bidang medis. Namun demikian, interpretasi manual terhadap citra rontgen membutuhkan tingkat keahlian tinggi dari radiolog, yang tidak selalu tersedia, terutama di wilayah terpencil di Indonesia. Oleh karena itu, teknologi AI menawarkan potensi besar untuk meningkatkan akurasi diagnosis melalui sistem deteksi otomatis berbasis citra [1]. Salah satu strategi yang berkembang adalah integrasi CNN dengan *vision transformer* (ViT) untuk memperkuat kemampuan analisis citra beresolusi tinggi [2].

Penelitian-penelitian sebelumnya telah memvalidasi kemampuan CNN dalam menangkap serta mengekstraksi pola spasial lokal pada data pencitraan medis secara efisien [3]. Di sisi lain, ViT memiliki kemampuan untuk memodelkan hubungan spasial global, sehingga memberikan pemahaman yang lebih mendalam dalam analisis citra rontgen [4]. Beberapa studi juga menunjukkan hasil yang menjanjikan melalui penerapan model yang awalnya dilatih, seperti EfficientNet dan ViT, yang kemudian disesuaikan (*fine-tuned*) untuk meningkatkan presisi deteksi pneumonia [5], [6]. Namun, sebagian besar model tersebut belum diimplementasikan pada konteks medis yang menuntut interpretasi hasil yang jelas.

Permasalahan lain dalam pengembangan AI untuk diagnosis pneumonia adalah kurangnya integrasi data multimodal. Sebagian besar penelitian hanya berfokus pada data visual dari citra rontgen tanpa mempertimbangkan data klinis tambahan yang berpotensi meningkatkan akurasi. Selain itu, penelitian klinis untuk menguji efektivitas model AI dalam



Gambar 1. Langkah-langkah pemrosesan citra yang diusulkan dalam studi ini.

lingkungan nyata masih terbatas. Hal ini menyebabkan adanya kesenjangan antara kinerja model di laboratorium dan penerapannya dalam praktik medis sehari-hari.

Penelitian ini bertujuan untuk mengembangkan model *explainable artificial intelligence* (XAI) berbasis CNN-ViT dengan visualisasi peta *saliency* pada proses inferensi. Model ini diharapkan dapat mendukung diagnosis pneumonia, khususnya di wilayah dengan keterbatasan sumber daya dan tenaga medis. Selain itu, penelitian ini juga mengintegrasikan informasi klinis nonvisual untuk meningkatkan ketepatan diagnosis serta relevansi klinis dari model yang diusulkan.

Penelitian-penelitian terdahulu menunjukkan bahwa kombinasi arsitektur hibrid CNN-ViT dengan metode interpretabilitas mampu meningkatkan reliabilitas dan efektivitas sistem diagnosis berbasis AI [6]. Studi ini juga mengintegrasikan data klinis di luar data pencitraan guna meningkatkan relevansi medis dan ketepatan diagnosis dari model yang dikembangkan [7]–[10]. Namun demikian, tantangan utama masih terletak pada pengembangan *dataset* global yang mencakup variasi demografis dan kondisi klinis serta pada pengujian efisiensi model. Penelitian ini berupaya menjembatani kesenjangan tersebut dengan berfokus pada kebutuhan klinis yang lebih mendalam, interpretasi hasil yang lebih baik, serta kompatibilitas model untuk diimplementasikan pada perangkat portabel. Penelitian ini diharapkan dapat memberikan kontribusi pada pengembangan sistem diagnosis pneumonia berbasis AI yang transparan dan dapat diintegrasikan ke dalam lingkungan klinis.

Model yang diusulkan mengombinasikan CNN dengan ViT untuk memodelkan ketergantungan spasial dalam citra rontgen dada. Berbeda dari metode berbasis CNN konvensional, pendekatan ini meningkatkan akurasi diagnosis dan transparansi model melalui penerapan *gradient-weighted class activation mapping* (Grad-CAM), yang secara visual mengidentifikasi area-area paling relevan yang memengaruhi prediksi model. Hasil utama dari penelitian ini mencakup pengurangan tingkat kesalahan klasifikasi *false positive* dan *false negative*, peningkatan kemampuan model dalam melakukan generalisasi terhadap data yang beragam, peningkatan interpretabilitas, serta optimisasi kerangka kerja ujung ke ujung yang komprehensif untuk deteksi pneumonia otomatis.

Penelitian ini juga memberikan manfaat sosial yang signifikan, seperti memungkinkan diagnosis yang lebih cepat

dan akurat, mengurangi beban kerja radiolog, memperluas akses terhadap teknologi diagnosis berbasis AI, serta membuka peluang penerapan untuk mendeteksi penyakit paru lainnya seperti tuberkulosis dan COVID-19. Dengan demikian, penelitian ini memberikan dasar yang kuat untuk penerapan AI secara lebih luas dalam bidang pencitraan medis, dengan menyediakan alat bantu diagnosis yang andal dan dapat dijelaskan.

## II. METODOLOGI

Penelitian ini melakukan deteksi pneumonia menggunakan citra rontgen dada berbasis pendekatan *deep learning*. Proses metode terdiri atas beberapa tahap utama, yaitu persiapan *dataset*, pengembangan arsitektur model CNN-ViT, evaluasi model, serta penerapan teknik *explainability* menggunakan Grad-CAM sebagaimana ditunjukkan pada Gambar 1. Setiap tahapan memiliki peran penting untuk memastikan model dapat beroperasi secara optimal dalam mendeteksi dan menginterpretasikan area-area kritis pada citra rontgen.

Sebelum merumuskan arsitektur hibrid CNN-ViT yang diusulkan dengan Grad-CAM, dilakukan serangkaian percobaan awal terhadap beberapa pendekatan alternatif. Pada tahap awal, arsitektur CNN konvensional seperti VGG16 dan ResNet50 diterapkan. Model-model tersebut menunjukkan akurasi yang memadai, tetapi terbatas dalam menangkap ketergantungan spasial global. Selanjutnya, dilakukan pengujian menggunakan ViT tunggal untuk mengatasi keterbatasan tersebut. Hasilnya menunjukkan penurunan kemampuan dalam mengekstraksi fitur lokal yang bersifat detail pada citra rontgen dada. Berdasarkan hasil percobaan awal tersebut, kombinasi CNN (untuk ekstraksi fitur lokal) dan ViT (untuk pemodelan konteks global) diusulkan, dengan integrasi Grad-CAM untuk meningkatkan interpretabilitas. Kombinasi ini memberikan keseimbangan terbaik antara akurasi diagnosis dan transparansi model.

Grad-CAM merupakan metode yang menghasilkan penjelasan visual terhadap keputusan model CNN dengan mengidentifikasi area citra yang berkontribusi signifikan terhadap keluaran model [8]. Metode ini memanfaatkan gradien yang mengalir ke lapisan konvolusional terakhir untuk menghasilkan peta aktivasi yang menyoroti area penting dalam citra yang memengaruhi prediksi model [9].

Grad-CAM juga telah diterapkan pada berbagai arsitektur model, termasuk ViT. Dalam ViT, keluaran setiap lapisan umumnya memiliki ukuran  $batch \times 197 \times 192$ . Elemen pertama

merepresentasikan token kelas, sedangkan elemen lainnya mewakili *patch*  $14 \times 14$  pada citra. Untuk menerapkan Grad-CAM pada ViT, aktivasi dan gradien dikonversi ke bentuk citra spasial 2D menggunakan fungsi *reshape\_transform*. Implementasi Grad-CAM untuk ViT tersedia dalam pustaka PyTorch-Grad-CAM, yang mendukung berbagai arsitektur model dan beragam skenario penggunaan [10], [11].

#### A. PERSIAPAN DATASET

Tahap persiapan *dataset* melibatkan beberapa langkah prapemrosesan untuk memastikan kualitas data yang optimal sebelum digunakan dalam pelatihan model. *Dataset* yang dipersiapkan dengan baik sangat penting untuk mencapai kinerja model yang tinggi dan meminimalkan bias yang dapat memengaruhi akurasi diagnosis.

Penelitian ini memanfaatkan *dataset* publik ChestX-ray8 dan RSNA [12]. *Dataset* ChestX-ray8 berisi citra rontgen dada tampak depan (*frontal-view*) yang dianotasi dengan 14 kategori penyakit toraks, termasuk pneumonia. Citra disediakan dalam format PNG dengan resolusi  $1.024 \times 1.024$  piksel. *Dataset* ini memuat citra pasien yang dikategorikan sebagai positif pneumonia dan normal. *Dataset* RSNA terdiri atas 26.684 citra rontgen dada dalam format DICOM, diperoleh dari pasien anak berusia antara 1 hingga 5 tahun. Setiap citra telah ditinjau dan dianotasi oleh dokter spesialis radiologi terkait keberadaan atau ketiadaan pneumonia.

*Dataset* dibagi secara cermat menjadi tiga bagian, yaitu data pelatihan, data validasi, dan data pengujian, untuk memastikan evaluasi kinerja model yang adil dan seimbang. Pembagian dilakukan dengan komposisi 1.926 citra untuk pelatihan, 1.306 citra untuk validasi, dan 573 citra untuk pengujian. Pembagian ini dirancang agar terdapat keseimbangan antara ketersediaan sampel untuk proses pembelajaran model dan jumlah data yang cukup untuk validasi serta pengujian yang tidak bias.

Beberapa metode prapemrosesan diterapkan untuk meningkatkan kemampuan generalisasi model, seperti augmentasi, normalisasi, dan penghilangan artefak. Tahapan ini membantu mengatasi permasalahan seperti ketidakseimbangan kelas, variasi kualitas citra, serta derau (*noise*) yang dapat menghambat proses pembelajaran.

##### 1) AUGMENTASI DATA

Ukuran *dataset* dalam pencitraan medis umumnya relatif kecil, sehingga terdapat risiko tinggi terjadinya *overfitting*, yaitu ketika model mengingat data pelatihan alih-alih mempelajari representasi yang dapat digeneralisasikan pada data baru [13]. Oleh karena itu, langkah pertama yang diterapkan adalah memperluas variasi *dataset* dan menggabungkan variasi merepresentasikan kondisi nyata.

Teknik augmentasi yang digunakan dalam penelitian ini meliputi rotasi acak citra dalam rentang  $\pm 15^\circ$  untuk mengantisipasi ketidaksejajaran ringan pada pengambilan rontgen; *horizontal flip* untuk menambah keragaman orientasi citra tanpa mengubah struktur anatomi; *random zoom* untuk mensimulasikan variasi tingkat pembesaran; serta penyesuaian kontras untuk meningkatkan ketahanan model terhadap perbedaan intensitas rontgen. Melalui penerapan teknik augmentasi ini, model dilatih untuk beradaptasi terhadap beragam kondisi citra pasien, sehingga meningkatkan kinerja pada aplikasi di dunia nyata.

##### 2) NORMALISASI

Standardisasi nilai piksel melalui proses normalisasi memiliki peran penting dalam memastikan distribusi masukan

yang konsisten, yang berdampak pada stabilitas pelatihan dan percepatan konvergensi model [14]. Semua citra rontgen dalam *dataset* dinormalisasi menggunakan nilai rata-rata dan simpangan baku yang telah ditentukan sebelumnya. Proses normalisasi dilakukan sesuai dengan (1).

$$X_{normalized} = \frac{X - \mu}{\sigma} \quad (1)$$

dengan  $X$  menunjukkan nilai piksel mentah,  $\mu$  adalah rata-rata *dataset*, dan  $\sigma$  adalah simpangan baku. Penyesuaian ini menormalkan nilai piksel agar terpusat di sekitar nol, sehingga mengurangi pengaruh perbedaan pencahayaan dan kontras antar citra.

Normalisasi juga membantu memastikan konsistensi antarsumber citra karena hasil rontgen dapat bervariasi, tergantung pada peralatan dan pengaturan *exposure* yang digunakan. Dengan melakukan standardisasi nilai piksel, model dapat lebih fokus pada pola yang relevan daripada perbedaan tingkat kecerahan atau kontras.

##### 3) PENGHILANGAN ARTEFAK DAN PENINGKATAN CITRA

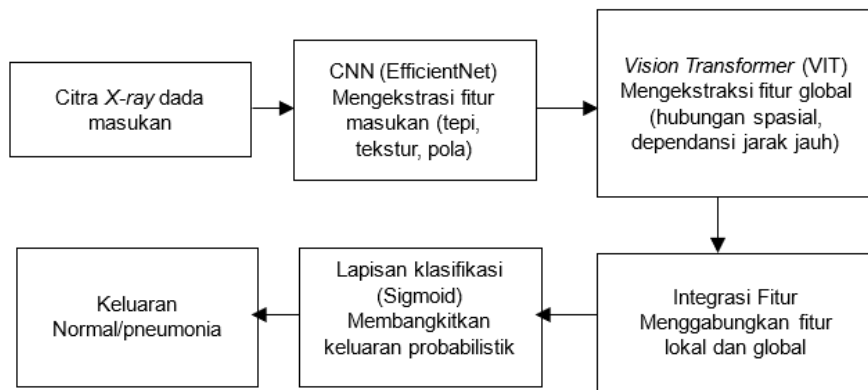
Citra rontgen sering kali mengandung artefak seperti label, derau, atau elemen latar belakang yang tidak relevan dan dapat menimbulkan bias dalam proses pembelajaran model [15]. Oleh karena itu, dilakukan beberapa tahapan prapemrosesan tambahan, termasuk pemangkasan (*cropping*) untuk menghapus tepi atau label yang tidak diperlukan; peningkatan kontras guna menonjolkan fitur anatomi penting untuk ekstraksi fitur yang lebih baik; serta penerapan filter Gaussian atau median untuk mengurangi derau sambil tetap mempertahankan struktur penting. Melalui penerapan teknik-teknik ini, *dataset* diperbaiki sehingga menonjolkan fitur medis yang relevan, yang pada akhirnya meningkatkan keandalan model AI dalam mendeteksi pneumonia.

#### B. PENGEMBANGAN ARSITEKTUR MODEL CNN-ViT

Diagram blok model CNN-ViT yang diusulkan ditunjukkan pada Gambar 2. Pada tahap ini, CNN dikombinasikan dengan ViT untuk meningkatkan kemampuan interpretasi citra medis. EfficientNet digunakan sebagai komponen CNN dasar yang berfungsi mengekstraksi fitur-fitur lokal secara detail dari citra rontgen dada. Model EfficientNet dikenal memiliki skalabilitas dan akurasi tinggi pada tugas klasifikasi citra. Arsitektur ini menerapkan metode *compound scaling*, yang secara proporsional meningkatkan kedalaman (*depth*), lebar (*width*), dan resolusi masukan jaringan. Pendekatan ini memungkinkan penangkapan karakteristik visual yang kompleks, seperti tepi, tekstur, dan pola yang umum muncul pada manifestasi pneumonia.

EfficientNet berperan sebagai tulang punggung (*backbone*) CNN untuk melakukan ekstraksi fitur lokal dari citra rontgen dada. Dengan kemampuan efisiensi dan ketepatan tinggi dalam klasifikasi citra, EfficientNet melakukan penskalaan gabungan pada kedalaman, lebar, dan resolusi jaringan secara proporsional, sehingga model dapat mengenali detail halus seperti struktur tepi, pola tekstur, serta ciri khas yang relevan terhadap deteksi pneumonia.

Untuk mengintegrasikan informasi kontekstual global, arsitektur ini juga mencakup ViT [16]. ViT memanfaatkan mekanisme *self-attention* untuk menangkap ketergantungan jarak jauh (*long-range dependencies*) dan hubungan spasial antararea dalam citra, yang sangat penting dalam menganalisis struktur anatomi kompleks serta variasi halus pada hasil pemindaian medis. Konteks global yang diperoleh melalui ViT



Gambar 2. Diagram blok pengembangan arsitektur model CNN-ViT.

melengkapi proses ekstraksi fitur lokal pada EfficientNet, sehingga menghasilkan representasi masukan yang lebih komprehensif.

Integrasi antara EfficientNet ViT memungkinkan model untuk memanfaatkan representasi fitur lokal dan global secara simultan guna meningkatkan kinerja keseluruhan. Modul CNN bertanggung jawab untuk mengekstraksi fitur lokal yang bersifat detail (*fine-grained*), sedangkan modul ViT berperan dalam menangkap ketergantungan spasial yang luas di seluruh area citra. Integrasi kedua modul ini memungkinkan model untuk mendeteksi dan mengklasifikasikan kelainan yang berkaitan dengan pneumonia pada citra rontgen dada dengan tingkat efektivitas yang lebih tinggi. Lapisan klasifikasi akhir menggunakan fungsi aktivasi *sigmoid* untuk menghasilkan keluaran probabilistik terhadap dua kelas, yaitu normal dan pneumonia.

Penelitian-penelitian terkini telah mengonfirmasi efektivitas penggabungan arsitektur CNN dan ViT dalam analisis citra medis. Sebuah penelaahan sistematis juga menyoroti bahwa integrasi ini mampu mengompensasi keterbatasan masing-masing jenis model, sehingga menghasilkan kerangka kerja yang seimbang antara kemampuan ekstraksi fitur lokal secara detail dan kesadaran konteks spasial yang lebih luas [17].

### C. EVALUASI MODEL

Model dilatih menggunakan algoritma optimasi Adam dengan laju pembelajaran awal sebesar 0,001 untuk memastikan proses penurunan gradien yang stabil serta mempercepat konvergensi. Ukuran *batch* ditetapkan sebesar 16 untuk menjaga keseimbangan antara beban komputasi dan efisiensi pelatihan. Untuk mengarahkan proses pembelajaran dan mengoptimalkan hasil klasifikasi, digunakan *loss function categorical cross-entropy*.

Pelatihan model dilakukan menggunakan Python 3 pada platform Google Colab, dengan memanfaatkan GPU A100 untuk mempercepat pemrosesan tugas komputasi yang intensif, seperti ekstraksi fitur dan pemodelan mekanisme *self-attention*. Durasi pelatihan dibatasi hingga lima *epoch* untuk memberikan waktu pembelajaran yang memadai tanpa menyebabkan *overfitting*.

Untuk mengevaluasi kinerja model secara komprehensif, digunakan beberapa metrik evaluasi, yaitu akurasi, presisi, *recall*, *confusion matrix*, serta *receiver operating characteristic* (ROC) dan *area under the curve* (AUC). Akurasi mengukur rasio prediksi benar terhadap total sampel, berfungsi sebagai indikator umum kinerja model. Presisi menggambarkan

proporsi prediksi positif yang benar dibandingkan seluruh prediksi positif, sedangkan *recall* (sensitivitas) menunjukkan kemampuan model dalam mendeteksi seluruh kasus positif yang sebenarnya. Kombinasi metrik-metrik tersebut memberikan pemahaman yang kompleks mengenai kinerja klasifikasi model. Selain itu, *confusion matrix* digunakan untuk memberikan analisis terperinci terhadap hasil klasifikasi, dengan menggambarkan jumlah sampel yang diklasifikasikan secara benar maupun salah pada seluruh kategori prediksi. *Confusion matrix* tersebut juga berkontribusi lebih lanjut dalam proses evaluasi dengan memberikan gambaran terperinci mengenai hasil prediksi, yaitu dengan membedakan antara *true positive*, *true negative*, *false positive*, dan *false negative*.

Untuk mengevaluasi keseimbangan antara sensitivitas dan spesifisitas pada berbagai ambang keputusan, digunakan kurva ROC. Nilai AUC yang bersesuaian mengukur kemampuan model secara keseluruhan dalam membedakan kedua kelas; makin tinggi nilai AUC, makin baik kemampuan diskriminatif model. Secara keseluruhan, metrik-metrik tersebut memberikan evaluasi menyeluruh terhadap kemampuan model dalam mengidentifikasi kasus normal dan pneumonia berdasarkan citra rontgen dada.

### D. IMPLEMENTASI EXPLAINABILITY (Grad-CAM)

Untuk memvisualisasikan proses pengambilan keputusan oleh model, diterapkan teknik Grad-CAM yang menghasilkan peta *saliency* untuk menunjukkan area pada citra masukan yang memiliki pengaruh paling signifikan terhadap keluaran model. Metode ini diimplementasikan pada lapisan konvolusi terakhir dari CNN, sehingga memungkinkan identifikasi fitur-fitur utama yang berkontribusi terhadap hasil prediksi model. *Heatmap* yang dihasilkan kemudian ditumpangkan (*superimposed*) pada citra rontgen dada asli, sehingga memberikan wawasan visual mengenai area paru-paru spesifik yang diidentifikasi oleh model sebagai indikator adanya pneumonia.

## III. HASIL DAN PEMBAHASAN

Sebelum proses pelatihan dilakukan, seluruh citra rontgen dada melalui tahapan prapemrosesan yang mencakup perubahan ukuran menjadi  $224 \times 224$  piksel, normalisasi, serta augmentasi data. Tahap prapemrosesan ini meningkatkan keseragaman citra, memperbaiki kontras, dan mengurangi variabilitas yang disebabkan oleh perbedaan kondisi akuisisi, sehingga model dapat belajar dari data masukan yang konsisten dan berkualitas tinggi. Langkah-langkah ini memberikan dasar

TABEL I  
TINJAUAN METRIK PELATIHAN MODEL CNN-ViT

Epoch	Loss	Akurasi Pelatihan (%)
Epoch 1	0,3670	89,01
Epoch 2	0,1325	94,93
Epoch 3	0,1003	96,09
Epoch 4	0,0946	96,46
Epoch 5	0,0870	96,89

yang andal bagi proses pelatihan model CNN-ViT yang diusulkan.

#### A. KINERJA PELATIHAN

Kinerja pelatihan dari model hibrid CNN-ViT dirangkum pada Tabel I, yang menunjukkan peningkatan akurasi dan penurunan nilai *loss* selama lima *epoch* pelatihan. Hasil tersebut mengindikasikan bahwa model menunjukkan proses pembelajaran yang stabil dan konsisten serta mampu mengekstraksi fitur-fitur bermakna dari *dataset* secara efektif.

Pada awal pelatihan (*epoch* 1), model mencapai akurasi sebesar 89,01%, yang menunjukkan bahwa model telah mulai mengenali pola-pola dasar dalam *dataset*. Namun, nilai *loss* yang relatif tinggi (0,3670) mengindikasikan masih adanya ruang untuk peningkatan. Seiring berjalannya pelatihan, model dengan cepat mengoptimalkan parameter-parameternya, menurunkan nilai *loss* menjadi 0,1325 pada *epoch* ke-2 dan secara signifikan meningkatkan akurasi hingga 94,93%.

Pada *epoch* ke-3, akurasi pelatihan melampaui 96%, sementara nilai *loss* menurun menjadi 0,1003, yang menunjukkan bahwa model makin mampu menyempurnakan representasi fiturnya. Kurva pelatihan menunjukkan kecenderungan stabil pada *epoch* berikutnya, dengan akurasi mencapai 96,89% pada *epoch* ke-5 dan nilai *loss* turun lebih lanjut menjadi 0,0870. Hasil ini mengonfirmasi bahwa model hampir mencapai konvergensi pada *epoch* terakhir, dengan kesalahan klasifikasi yang berhasil diminimalkan secara efektif.

Penurunan nilai *loss function* yang konsisten, sebagaimana ditunjukkan pada Tabel I, merupakan indikator kuat bahwa model berhasil mempelajari pola-pola dalam *dataset* sambil menghindari permasalahan utama, seperti divergensi atau ketidakstabilan selama proses pelatihan. Integrasi antara CNN untuk ekstraksi fitur spasial lokal dan ViT untuk pemodelan hubungan kontekstual global secara signifikan meningkatkan efektivitas pembelajaran model.

Namun, potensi terjadinya *overfitting* perlu dipantau karena akurasi pelatihan yang sangat tinggi dapat mengindikasikan bahwa model cenderung menghafal pola pada data pelatihan, alih-alih melakukan generalisasi secara efektif. Kekhawatiran ini dibahas lebih lanjut pada bagian berikut. Kinerja model dievaluasi menggunakan data uji yang belum pernah dilihat sebelumnya (*unseen test data*).

Model dilatih selama lima *epoch* untuk mencegah *overfitting* yang mungkin terjadi akibat ukuran *dataset* yang relatif kecil setelah melalui tahap prapemrosesan dan augmentasi. Hasil percobaan awal menunjukkan bahwa kinerja model mulai mencapai kondisi *plateau* setelah *epoch* ke-4, dengan peningkatan akurasi yang minimal serta peningkatan risiko *overfitting* jika pelatihan dilanjutkan melebihi lima *epoch*.

TABEL II  
RINGKASAN MATRIKS KLASIFIKASI UNTUK ARSITEKTUR CNN-ViT

	Precision	Recall	Skor F1	Support
Normal	0,90	0,98	0,942	148
Pneumonia	0,99	0,94	0,98	425
Akurasi			0,96	573
Rerata makro	0,96	0,96	0,93	573
Weighted average	0,95	0,976	0,94	573

#### B. EVALUASI MODEL PADA DATASET UJI

Kinerja model pada *dataset* uji dijabarkan dalam Tabel II, yang menyajikan nilai presisi, *recall*, dan skor F1 untuk kelas normal dan pneumonia. Model mencapai akurasi klasifikasi yang tinggi sebesar 97%, dengan skor F1 sebesar 0,98 untuk deteksi pneumonia dan 0,94 untuk kasus normal. Hasil ini menunjukkan kemampuan model yang kuat dalam membedakan kasus pneumonia dari kasus normal

Untuk memvalidasi kemampuan generalisasi model, digunakan *dataset* uji terpisah guna memastikan bahwa pola-pola yang telah dipelajari selama tahap pelatihan dapat diterapkan secara efektif pada data yang belum pernah dilihat sebelumnya. Metrik kinerja klasifikasi yang disajikan dalam Tabel II menegaskan ketangguhan model terhadap berbagai kriteria evaluasi utama, termasuk presisi, *recall*, dan skor F1 untuk kedua kelas tersebut.

Model menunjukkan akurasi keseluruhan yang mengesankan, sebesar 97%, yang mencerminkan kemampuan klasifikasi yang kuat. Nilai presisi dan *recall* untuk masing-masing kelas makin memperkuat kesimpulan ini.

- Untuk kasus-kasus normal, model mencapai presisi sebesar 0,91 dan *recall* sebesar 0,97 serta menghasilkan skor F1 sebesar 0,94. Hasil ini menunjukkan bahwa meskipun model sedikit konservatif dalam memprediksi kasus normal (terlihat dari nilai presisi), model tetap berhasil mengenali sebagian besar kasus normal yang sebenarnya (nilai *recall* tinggi).
- Untuk kasus-kasus pneumonia, model mencapai presisi sebesar 0,99 dan *recall* sebesar 0,96, dengan skor F1 sebesar 0,98. Hasil ini menunjukkan bahwa model sangat efektif dalam mengklasifikasikan kasus pneumonia secara benar, dengan tingkat kesalahan *false negative* yang sangat rendah.

Nilai rerata makro (rata-rata dari presisi, *recall*, dan skor F1 pada kedua kelas) mencapai 0,96, sedangkan *weighted average* (yang mempertimbangkan ketidakseimbangan jumlah sampel antarkelas) sebesar 0,97. Nilai-nilai ini menunjukkan bahwa model memiliki kinerja yang baik pada kedua kategori, tanpa menunjukkan kecenderungan berlebih terhadap salah satu kelas.

Nilai *recall* yang tinggi pada kasus pneumonia memiliki signifikansi penting dalam konteks diagnosis medis karena kegagalan dalam mendeteksi satu kasus pneumonia (*false negative*) dapat menimbulkan konsekuensi serius. Sementara itu, nilai presisi yang tinggi pada kasus pneumonia juga berperan dalam meminimalkan *false positive*, sehingga mengurangi risiko intervensi yang tidak diperlukan maupun kesalahan diagnosis.

Nilai *recall* yang tinggi pada kasus pneumonia memiliki signifikansi penting dalam konteks diagnosis medis karena kegagalan dalam mendeteksi satu kasus pneumonia (*false*

*negative*) dapat menimbulkan konsekuensi serius. Sementara itu, nilai presisi yang tinggi pada kasus pneumonia juga berperan dalam meminimalkan *false positive*, sehingga mengurangi risiko intervensi yang tidak diperlukan maupun kesalahan diagnosis.

Untuk makin memastikan keandalan model, dilakukan analisis terhadap *confusion matrix* dan nilai ROC–AUC. Nilai ROC–AUC yang ideal adalah mendekati 1,0, mencerminkan kemampuan diskriminatif model yang tinggi dalam membedakan antara kasus normal dan pneumonia. Hasil analisis menunjukkan bahwa arsitektur hibrid CNN–ViT berhasil mencapai keseimbangan yang efektif antara sensitivitas dan spesifisitas, yang mengindikasikan keandalannya sebagai alat bantu diagnosis pneumonia berbasis citra rontgen dada. Kinerja kuat ini dapat dikaitkan dengan integrasi antara ekstraksi fitur lokal berbasis CNN dan mekanisme *global attention* yang digerakkan oleh ViT, sejalan dengan penelitian terkini yang mendukung efektivitas model *deep learning* hibrid dalam aplikasi pencitraan medis.

Ambang probabilitas klasifikasi sebesar 0,5 diterapkan untuk membedakan antara kasus normal dan pneumonia. Untuk keperluan perbandingan kinerja (*performance benchmarking*), model yang diusulkan dibandingkan dengan model dasar (*baseline models*), yaitu VGG16, ResNet50, dan ViT tunggal, yang seluruhnya dilatih dan dievaluasi dalam kondisi yang identik. Hasil perbandingan disajikan pada Tabel III.

Gambar 3 menunjukkan *confusion matrix* yang menggambarkan kinerja model dalam mengklasifikasikan kasus pneumonia berdasarkan citra rontgen dada. Matriks tersebut memperlihatkan bahwa model mencapai hasil klasifikasi yang baik, dengan proporsi prediksi benar yang tinggi. Pada kelas normal, model berhasil mengklasifikasikan 139 sampel secara benar, tetapi 9 sampel salah diklasifikasikan sebagai pneumonia. Sementara itu, pada kelas pneumonia, model mampu mengklasifikasikan 423 sampel dengan benar, tetapi 2 sampel salah diklasifikasikan sebagai normal.

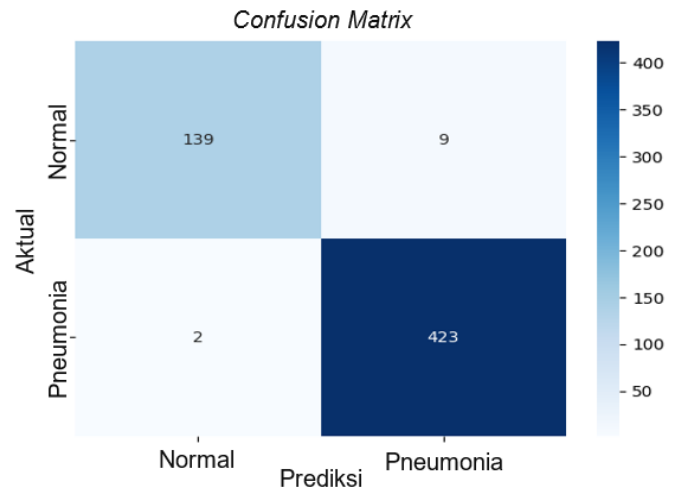
Seperti ditunjukkan pada Gambar 3, model memperlihatkan tingkat kesalahan yang sangat rendah, khususnya pada kasus *false negative*, dengan hanya dua kasus pneumonia yang salah diklasifikasikan sebagai normal. Meskipun jumlahnya minimal, kesalahan tersebut tetap memiliki signifikansi klinis karena kegagalan dalam mendeteksi pneumonia dapat menimbulkan konsekuensi kesehatan yang serius. Secara keseluruhan, hasil ini menunjukkan bahwa model mencapai tingkat akurasi diagnostik yang tinggi dan memiliki potensi untuk diterapkan dalam sistem diagnosis berbantuan AI di praktik radiologi.

Hasil evaluasi kinerja menunjukkan bahwa model hibrid CNN–ViT mencapai akurasi sebesar 96,5%, dengan nilai presisi, *recall*, dan skor F1 yang tinggi pada kedua kategori, yakni pneumonia dan normal. *Confusion matrix* (Gambar 3) makin menegaskan keandalan model dalam membedakan kedua kondisi tersebut, dengan tingkat prediksi salah yang sangat rendah.

Validasi lebih lanjut ditunjukkan melalui kurva ROC pada Gambar 4, yang memberikan penilaian terperinci terhadap kemampuan diskriminatif model. Dengan nilai AUC sebesar 1,00, model menunjukkan kemampuan hampir sempurna dalam membedakan antara kasus normal dan pneumonia serta menunjukkan sensitivitas dan spesifisitas yang sama-sama tinggi.

TABEL III  
PERBANDINGAN MODEL USULAN DENGAN MODEL-MODEL DASAR

Model	Akurasi (%)	Presisi (%)	Recall (%)	Skor F1 (%)
VGG16	92.4	91.8	92.6	92.2
ResNet50	93.5	93.1	93.8	93.4
ViT	94.1	93.9	94.5	94.2
Metode usulan	96.5	96	96	94



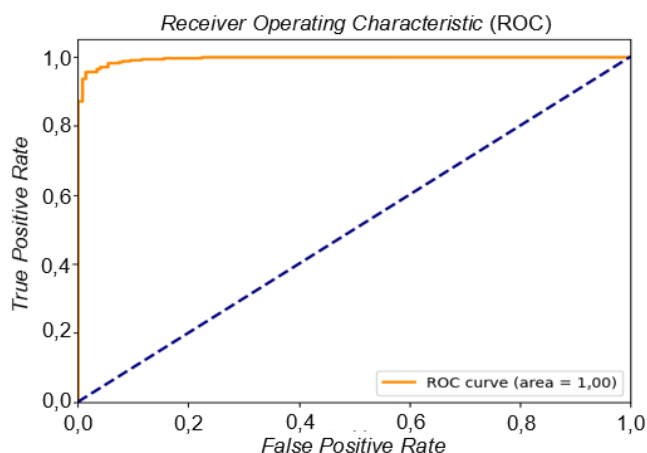
Gambar 3. *Confusion matrix* model CNN–ViT model, menunjukkan klasifikasi yang akurat untuk kasus normal dan pneumonia, dengan tingkat *true positive* dan *true negative* yang tinggi.

Kurva ROC yang melakukan *plot* rasio *true positive* (sensitivitas) terhadap rasio *false positive* ( $1 - \text{specificity}$ ) menggambarkan kemampuan model dalam memaksimalkan deteksi yang benar sekaligus meminimalkan peringatan yang keliru. Kenaikan tajam di dekat titik asal serta *plateau* di sudut kiri atas *plot* menunjukkan kinerja *recall* yang sangat baik dengan tingkat *false positive* yang minimal. Kondisi ini sangat menguntungkan dalam konteks diagnosis medis karena memastikan tidak adanya kasus pneumonia yang terlewat merupakan hal yang krusial bagi keselamatan pasien dan efektivitas pengobatan.

Nilai AUC yang mendekati sempurna menunjukkan bahwa model hibrid CNN–ViT secara efektif memanfaatkan kombinasi ekstraksi fitur lokal dari CNN dan pemodelan hubungan spasial global dari ViT, sehingga sangat andal untuk aplikasi klinis. Penerapan mekanisme *self-attention* dalam ViT memungkinkan model mengenali pola kontekstual global pada citra medis, meningkatkan ketangguhan terhadap variasi data masukan, sekaligus memperkuat kemampuan prediksi yang dapat dijelaskan (*explainable predictions*).

Selain itu, hasil tersebut menunjukkan korelasi yang kuat dengan akurasi klasifikasi tinggi yang telah diamati sebelumnya (97%), nilai presisi sebesar 0,99 untuk kasus pneumonia, dan *recall* sebesar 0,96 untuk kasus pneumonia, yang makin menegaskan kemampuan generalisasi model terhadap berbagai kondisi. Tingkat kinerja ini menyoroti potensi signifikan untuk penerapan di dunia nyata, yaitu alat radiologi berbasis AI dapat membantu tenaga medis dalam mendeteksi pneumonia secara cepat dan akurat.

Lebih jauh lagi, penggunaan Grad-CAM meningkatkan interpretabilitas model dengan menyoroti secara visual area



**Gambar 4.** Kurva ROC menunjukkan klasifikasi model dengan nilai AUC sebesar 1,00, yang mengindikasikan kemampuan sangat baik dalam membedakan antara kedua kelas.

spesifik pada citra rontgen dada yang berkontribusi signifikan terhadap hasil prediksi. Lapisan *explainability* ini berperan penting dalam membangun kepercayaan klinis, karena memungkinkan praktisi kesehatan menilai dan memverifikasi dasar rasional di balik hasil diagnosis yang dihasilkan oleh AI.

Gambar 5 menampilkan keluaran Grad-CAM yang memperlihatkan kemampuan interpretatif model saat memproses citra rontgen dada normal. Metode visualisasi ini menekankan area citra yang paling relevan terhadap keputusan klasifikasi model. Gambar tersebut terdiri atas tiga panel.

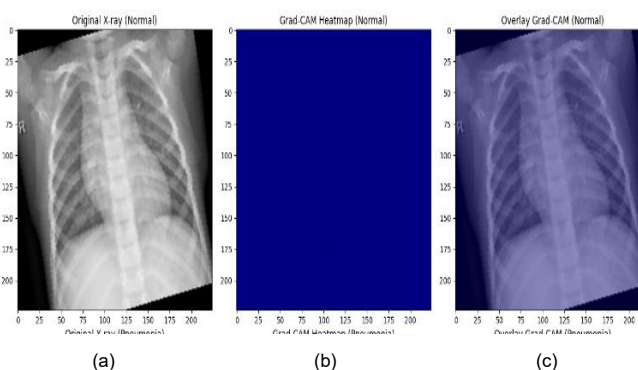
Gambar 5(a) menunjukkan citra rontgen dada asli tanpa modifikasi dari pasien tanpa pneumonia. Citra ini berfungsi sebagai masukan ke model CNN-ViT dan menampilkan struktur anatomi seperti medan paru dan tulang rusuk tanpa adanya kelegapan atau kelainan yang umumnya menunjukkan infeksi paru.

Gambar 5(b) merepresentasikan peta aktivasi yang dihasilkan oleh Grad-CAM, yang menyoroti area yang dianggap signifikan oleh model untuk proses klasifikasi. Pada kasus ini, *heatmap* hampir seluruhnya berwarna biru tua, menandakan bahwa model tidak mendeteksi adanya fitur abnormal pada citra tersebut. Tidak adanya area yang disorot mengonfirmasi bahwa model mengklasifikasikan kasus ini sebagai normal secara tepat tanpa salah mengenali area yang tidak relevan sebagai indikasi pneumonia.

Visualisasi pada Gambar 5(c) menampilkan *heatmap* Grad-CAM yang ditumpangkan (*overlaid*) pada citra rontgen asli. Tidak terlihat adanya area merah atau kuning yang menonjol, yang menunjukkan bahwa model tidak menemukan area mencurigakan yang memerlukan perhatian. Tumpang tindih yang bersih ini menunjukkan bahwa model dapat membedakan struktur paru normal dengan yakin tanpa salah mengidentifikasinya sebagai temuan patologis.

Temuan ini menunjukkan bahwa model hibrid CNN-ViT memiliki spesifisitas tinggi dalam mengidentifikasi kasus normal secara akurat, sehingga mampu meminimalkan prediksi *false positive* yang dapat menyebabkan prosedur klinis atau intervensi yang tidak diperlukan. Ketergantungan model pada fitur spasial dan kontekstual yang bermakna memastikan bahwa fokus klasifikasi diarahkan pada pola medis yang relevan, bukan pada derau latar atau variasi anatomi yang tidak signifikan.

Selain itu, efektivitas Grad-CAM sebagai alat *explainability* terlihat jelas melalui hasil visualisasi ini. Dalam



**Gambar 5.** Perbandingan citra rontgen dan peta Grad-CAM yang dihasilkan pada kelas normal, (a) citra rontgen asli, (b) Grad-CAM *Heatmap*, (c) *overlay* Grad-CAM.

aplikasi klinis, *interpretability* merupakan aspek penting bagi model diagnosis berbasis AI karena memungkinkan radiolog untuk memverifikasi proses pengambilan keputusan. Tidak adanya aktivasi *heatmap* yang menunjukkan kesalahan klasifikasi memperkuat tingkat kepercayaan dan keandalan keluaran sistem, yang menjadi faktor krusial dalam penerapan AI pada alur kerja pencitraan medis.

Secara keseluruhan, visualisasi ini menegaskan ketangguhan model CNN-ViT, yang mencapai akurasi klasifikasi tinggi dan mempertahankan batas keputusan yang jelas, sehingga mampu membedakan citra rontgen normal dan abnormal dengan risiko salah klasifikasi yang minimal. Aplikasi lanjutan dari model ini berpotensi diterapkan dalam sistem pendukung keputusan *real-time* bagi radiolog, guna meningkatkan efisiensi dan akurasi diagnosis pneumonia.

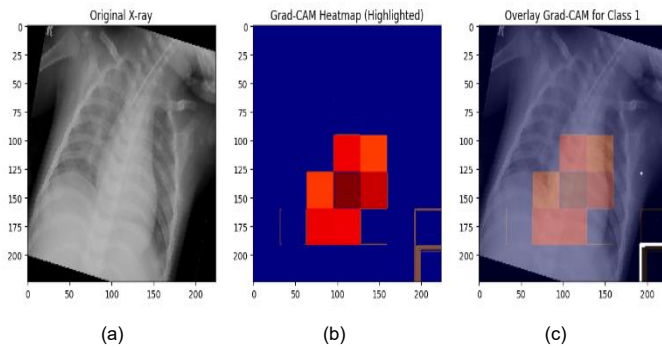
Keluaran Grad-CAM yang ditampilkan pada Gambar 6 memberikan interpretasi transparan mengenai cara model hibrid CNN-ViT menghasilkan keputusan diagnosis pneumonia berdasarkan citra rontgen dada. Gambar tersebut terdiri atas tiga elemen utama.

Gambar 6(a) menampilkan citra rontgen dada mentah dari pasien yang terdiagnosis pneumonia. Dibandingkan dengan paru normal, beberapa area paru tampak lebih legap, yang sering kali menandakan adanya akumulasi cairan atau peradangan—ciri khas umum dari pneumonia.

Gambar 6(b) menunjukkan *heatmap* yang dihasilkan oleh Grad-CAM, yang menyoroti area paling relevan dalam proses pengambilan keputusan klasifikasi. Pada kasus ini, *heatmap* memperlihatkan area berintensitas tinggi berwarna merah dan jingga, terutama pada bagian bawah dan tengah medan paru. Area tersebut merepresentasikan lokasi tempat model mendeteksi fitur abnormal, seperti opasitas atau pola yang konsisten dengan karakteristik pneumonia.

Gambar 6(c) memperlihatkan *heatmap* yang ditindihkan pada citra rontgen asli, memberikan representasi visual terhadap mekanisme *attention model*. Area merah terang menunjukkan aktivasi yang kuat, yang berarti model berfokus pada area tersebut saat menentukan keberadaan pneumonia. Lokalisasi aktivasi yang presisi ini mengindikasikan bahwa model secara efektif mengenali area yang relevan secara klinis tanpa memusatkan perhatian pada area yang tidak signifikan, sehingga meningkatkan keandalan diagnosis.

Hasil visualisasi Grad-CAM ini mengonfirmasi bahwa model tidak hanya akurat, tetapi juga dapat dijelaskan (*explainable*), sehingga meningkatkan *interpretability* dalam sistem diagnosis medis berbasis AI. Kemampuan untuk memvisualisasikan proses pengambilan keputusan sangat



**Gambar 6.** Perbandingan citra rontgen dan peta Grad-CAM yang dihasilkan pada kelas pneumonia. (a) Citra rontgen asli, (b) *Heatmap* Grad-CAM, (c) *Overlay* Grad-CAM.

penting dalam aplikasi klinis karena memungkinkan radiolog memverifikasi bahwa model AI mengidentifikasi area paru yang benar-benar terdampak pneumonia, bukan membuat prediksi secara acak.

Selain itu, visualisasi *heatmap* Grad-CAM menunjukkan pola aktivasi yang terlokalisasi, menandakan bahwa model hibrid CNN–ViT berhasil menangkap fitur lokal secara detail melalui komponen CNN dan ketergantungan spasial yang lebih luas melalui ViT. Kemampuan ganda ini berkontribusi pada kinerja model yang tangguh dalam mendeteksi pneumonia secara akurat. Sinergi antara mekanisme ekstraksi fitur konvolusional dan *self-attention* meningkatkan sensitivitas dan spesifisitas model dalam mengenali kelainan yang terkait dengan pneumonia.

Dalam konteks klinis, interpretabilitas seperti ini sangat penting untuk membangun kepercayaan terhadap sistem diagnosis berbantuan AI. Integrasi Grad-CAM sebagai alat penjelasan visual memberikan transparansi terhadap proses pengambilan keputusan model, sehingga meningkatkan keandalan serta nilai praktisnya dalam evaluasi radiologis.

Secara khusus, hasil Grad-CAM menunjukkan bahwa model berfokus pada area tengah dan atas paru—bagian yang umumnya terpengaruh oleh pneumonia. Pola aktivasi mencakup area yang lebih luas dibandingkan pada kasus normal, sejalan dengan manifestasi klinis pneumonia sebagai infeksi paru difus. *Overlay heatmap* juga menunjukkan gradasi warna yang lebih intens pada area yang lebih besar, mengindikasikan bahwa model mengidentifikasi kelainan dengan tingkat keyakinan dan relevansi diagnostik yang tinggi. menunjukkan peningkatan kinerja, baik dari segi akurasi maupun interpretabilitas. Referensi [18] menggunakan CNN konvensional dan mencapai akurasi sebesar 93%, sedangkan pendekatan berbasis *transformer* dalam penelitian ini berhasil meningkatkan akurasi hingga 96,5%. Studi lain menekankan pentingnya aspek interpretabilitas pada model AI di bidang medis, yang dalam penelitian ini diakomodasi melalui penerapan Grad-CAM [19]. Keunggulan utama model yang dikembangkan terletak pada kemampuannya menghasilkan *excellence map* yang membantu dokter dalam memahami hasil diagnosis. Selain itu, integrasi model *transformer* memungkinkan deteksi fitur dengan cakupan yang lebih luas dibandingkan pendekatan CNN murni. Namun, tantangan yang masih dihadapi adalah kebutuhan daya komputasi yang lebih tinggi dibandingkan dengan model CNN konvensional. Dengan hasil tersebut, model hibrid CNN–ViT yang dikembangkan berpotensi diimplementasikan dalam lingkungan medis nyata, khususnya di daerah dengan

keterbatasan sumber daya, sekaligus tetap memberikan akurasi tinggi serta interpretasi visual yang andal.

Visualisasi Grad-CAM yang dihasilkan dalam penelitian ini menunjukkan efektivitas model dalam mengidentifikasi pneumonia pada citra rontgen dada. Hasilnya menunjukkan bahwa model usulan mampu membedakan secara akurat antara kasus normal dan pneumonia tanpa menyoroti area yang tidak relevan pada citra individu sehat. Selain itu, penggunaan Grad-CAM memfasilitasi interpretasi visual yang jelas dengan menunjukkan area spesifik pada paru-paru yang memengaruhi prediksi model, sehingga mendukung validasi klinis serta meningkatkan transparansi model dalam konteks diagnosis.

Hasil visualisasi Grad-CAM yang diperoleh juga menunjukkan bahwa model yang dikembangkan mampu membedakan citra rontgen normal dan pneumonia dengan tingkat akurasi yang baik tanpa menandai area yang tidak relevan pada pasien sehat. Pada citra rontgen normal, peta Grad-CAM tidak menunjukkan aktivasi signifikan, yang menandakan bahwa model tidak salah mengklasifikasikan area paru sebagai pneumonia. Sementara itu, pada citra rontgen dengan pneumonia, Grad-CAM berhasil menyoroti area paru bagian tengah dan atas, sesuai dengan karakteristik pneumonia yang umumnya menyebar pada jaringan paru. Sebagai contoh, sebuah studi pada *dataset* citra rontgen COVID-19 menggunakan Grad-CAM menunjukkan bahwa model *deep learning* mampu mengidentifikasi area abnormal dengan tingkat akurasi tinggi, meskipun pada beberapa kasus aktivasi menyebar ke area paru yang lebih luas [20]. Selain itu, sebuah penelitian yang menggunakan COVID-Net melaporkan bahwa Grad-CAM dapat digunakan untuk membantu dokter memahami cara model AI mengklasifikasikan penyakit paru [21]. Namun, penelitian lain menunjukkan keterbatasan dalam interpretasi Grad-CAM, yaitu model terkadang menangkap area di luar paru yang tidak relevan terhadap klasifikasi penyakit [22].

Temuan ini konsisten dengan penelitian-penelitian sebelumnya yang memanfaatkan Grad-CAM untuk menginterpretasikan model *deep learning* dalam deteksi penyakit paru. Sebagai contoh, sebuah studi mengusulkan arsitektur hibrid yang menggabungkan ResNet-50 (sebuah CNN) dengan ViT-b16 untuk mendeteksi tuberkulosis dan membedakannya dari pneumonia [7]. Model tersebut menunjukkan akurasi deteksi tuberkulosis sebesar 98,97% dan mencapai akurasi 96,18% dalam tugas klasifikasi multikelas, menegaskan potensinya dalam meningkatkan ketepatan diagnosis kondisi toraks melalui analisis citra rontgen dada.

Sebuah penelitian lain memperkenalkan arsitektur gabungan yang memanfaatkan ViT dan CNN untuk melakukan klasifikasi multikelas dan multilabel terhadap kelainan yang berkaitan dengan tuberkulosis pada citra rontgen dada [23]. Dengan tingkat akurasi 91,1%, model tersebut menunjukkan potensi yang kuat dalam meningkatkan deteksi kelainan yang berhubungan dengan tuberkulosis.

Penelitian lain juga mengintegrasikan ViT dan CNN pada pemindaian MRI otak [24]. Selain meningkatkan akurasi segmentasi, model tersebut menghasilkan *heatmap* yang dapat diinterpretasikan untuk membantu pengambilan keputusan pembedahan, sehingga meningkatkan transparansi dan menumbuhkan kepercayaan yang lebih besar dalam penerapan klinis. Penelitian lainnya mengusulkan model hibrid yang menggabungkan *Deep CNN* (DCNN) dengan ViT untuk mendukung diagnosis berbagai kondisi paru, termasuk kanker

paru, melalui pencitraan medis [25]. Pendekatan terintegrasi ini memanfaatkan kombinasi antara ekstraksi fitur lokal dan pemodelan konteks global untuk meningkatkan akurasi diagnosis pada berbagai kelainan toraks. Model ini menunjukkan peningkatan dalam representasi fitur dan akurasi diagnostik.

Dibandingkan dengan penelitian-penelitian sebelumnya, hasil penelitian ini menunjukkan bahwa model yang digunakan memiliki keunggulan dalam mempertahankan spesifisitas tinggi, terutama dalam meminimalkan *false positive* pada citra rontgen normal. Selain itu, metode ini juga mampu menyoroti area paru yang lebih terfokus dibandingkan dengan beberapa penelitian yang melaporkan pola aktivasi lebih menyebar. Oleh karena itu, penelitian ini mendukung penggunaan Grad-CAM sebagai alat interpretasi dalam model *deep learning* untuk deteksi pneumonia dan menunjukkan bahwa metode ini dapat digunakan untuk meningkatkan transparansi model AI di bidang radiologi.

#### IV. KESIMPULAN

Model CNN-ViT yang diusulkan secara efektif menggabungkan ekstraksi fitur lokal dengan pemahaman konteks global, menghasilkan peningkatan kinerja dalam deteksi pneumonia menggunakan citra rontgen dada. Hasil eksperimen menunjukkan bahwa model mencapai akurasi sebesar 96,5%, presisi sebesar 96%, *recall* sebesar 96%, dan skor F1 sebesar 94%, yang memvalidasi ketangguhan dan keandalannya. Temuan ini menyoroti potensi arsitektur hibrid CNN-Transformer dalam memajukan sistem diagnosis berbantuan komputer untuk aplikasi pencitraan medis.

Namun demikian, masih terdapat beberapa tantangan yang perlu diatasi, seperti kebutuhan daya komputasi yang tinggi serta potensi bias dalam interpretasi Grad-CAM. Oleh karena itu, penelitian selanjutnya dapat difokuskan pada optimisasi model agar lebih ringan dan dapat diimplementasikan pada perangkat medis berbasis *edge computing* serta pada integrasi data multimodal untuk meningkatkan relevansi medis. Penelitian di masa depan juga dapat mengeksplorasi *dataset* yang lebih besar dan beragam guna memvalidasi serta memperluas generalisasi efektivitas model.

#### KONFLIK KEPENTINGAN

Para penulis menyatakan bahwa tidak terdapat konflik kepentingan.

#### KONTRIBUSI PENULIS

Konseptualisasi, Atika Hendryani; metodologi, Atika Hendryani dan Vita Nurdinawati; perangkat lunak, Atika Hendryani; validasi, Vita Nurdinawati dan Agus Komarudin; analisis formal, Agus Komarudin; penulisan draf asli, Atika Hendryani, Vita Nurdinawati, dan Agus Komarudin.

#### REFERENSI

- [1] S.M. Shafi dan S.K. Chinnappan, "Hybrid transformer-CNN and LSTM model for lung disease segmentation and classification," *PeerJ. Comput. Sci.*, vol. 10, hal. 1–57, Des. 2024, doi: 10.7717/peerj-cs.2444.
- [2] C.C. Ukwuoma dkk., "A hybrid explainable ensemble transformer encoder for pneumonia identification from chest X-ray images," *J. Adv. Res.*, vol. 48, hal. 191–211, Jun. 2023, doi: 10.1016/j.jare.2022.08.021.
- [3] F.A. Mostafa, L.A. Elrefaei, M.M. Fouda, dan A. Hossam, "A survey on AI techniques for thoracic diseases diagnosis using medical images," *Diagnostics*, vol. 12, no. 12, hal. 1–60, Des. 2022, doi: 10.3390/DIAGNOSTICS12123034.
- [4] D. Pantelaios, P.A. Theofilou, P. Tzouveli, dan S. Kollias, "Hybrid CNN-ViT models for medical image classification," dalam *2024 IEEE Int.*

- Symp. Biomed. Imaging (ISBI)*, 2024, hal. 1–4, doi: 10.1109/ISBI56570.2024.10635205.
- [5] S.A. El-Ghany, M. Elmogy, M.A. Mahmood, dan A.A. Abd El-Aziz, "A robust tuberculosis diagnosis using chest X-rays based on a hybrid vision transformer and principal component analysis," *Diagnostics*, vol. 14, no. 23, hal. 1–36, Des. 2024, doi: 10.3390/DIAGNOSTICS14232736.
- [6] A.I. Kaleel dan S.B. Rajakumari, "Hybrid CNN-ViT model with Grad-CAM for ocular disease detection," dalam *2025 Int. Conf. Front. Technol. Solut. (ICFTS)*, 2025, hal. 1–7, doi: 10.1109/icfts62006.2025.11031986.
- [7] Y. Hadhoud dkk., "From binary to multi-class classification: A two-step hybrid CNN-ViT model for chest disease classification based on X-Ray images," *Diagnostics*, vol. 14, no. 23, hal. 1–16, Des. 2024, doi: 10.3390/DIAGNOSTICS14232754.
- [8] D. Bhati, F. Neha, dan M. Amiruzzaman, "A survey on explainable artificial intelligence (XAI) techniques for visualizing deep learning models in medical imaging," *J. Imaging*, vol. 10, no. 10, hal. 1–26, Okt. 2024, doi: 10.3390/JIMAGING10100239.
- [9] T. Kittani dan A.M.A. Albrifkani, "Deep learning classification algorithms applications: A review," *Indones. J. Comput. Sci.*, vol. 13, no. 3, hal. 1–25, Jun. 2024, doi: 10.33022/ijcs.v13i3.4064.
- [10] R.J.M. Veiga dan J.M.F. Rodrigues, "Fine-grained fish classification from small to large datasets with vision transformers," *IEEE Access*, vol. 12, hal. 113642–113660, 2024, doi: 10.1109/ACCESS.2024.3443654.
- [11] G.M.S. Himel dan M.M. Islam, "Benchmark analysis of various pre-trained deep learning models on ASSIRA cats and dogs dataset," *J. Umm Al-Qura Univ. Eng. Archit.*, vol. 16, no. 1, hal. 134–149, Mar. 2025, doi: 10.1007/S43995-024-00094-W.
- [12] D. J. Klionsky dkk., "Guidelines for the use and interpretation of assays for monitoring autophagy (3rd edition)," *Autophagy*, vol. 12, no. 1, hal. 1–222, 2016, doi: 10.1080/15548627.2015.1100356.
- [13] P. Chlap dkk., "A review of medical image data augmentation techniques for deep learning applications," *J. Med. Imaging Radiat. Oncol.*, vol. 65, no. 5, hal. 545–563, Agu. 2021, doi: 10.1111/1754-9485.13261.
- [14] A. Carré dkk., "Standardization of brain MR images across machines and protocols: Bridging the gap for MRI-based radiomics," *Sci. Rep.*, vol. 10, hal. 1–15, Jul. 2020, doi: 10.1038/s41598-020-69298-z.
- [15] F. Pérez-García, R. Sparks, dan S. Ourselin, "TorchIO: A Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning," *Comput. Methods Programs Biomed.*, vol. 208, hal. 1–12, Sep. 2021, doi: 10.1016/j.cmpb.2021.106236.
- [16] E.U. Henry, O. Emebob, dan C.A. Omonhinmin, "Vision transformers in medical imaging: A review," 2022, *arXiv:2211.10043*.
- [17] J.W. Kim, A.U. Khan, dan I. Banerjee, "Systematic review of hybrid vision transformer architectures for radiological image analysis," *J. Imaging Inform. Med.*, hal. 1–15, Jan. 2025, doi: 10.1007/S10278-024-01322-4.
- [18] T. Wang dkk., "PneuNet: Deep learning for COVID-19 pneumonia diagnosis on chest X-ray image analysis using vision transformer," *Med. Biol. Eng. Comput.*, vol. 61, no. 6, hal. 1395–1408, Jun. 2023, doi: 10.1007/S11517-022-02746-2/TABLES/10.
- [19] Y. Yang, G. Mei, dan F. Piccialli, "A deep learning approach considering image background for pneumonia identification using explainable AI (XAI)," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 21, no. 4, hal. 857–868, Jul./Agu. 2024, doi: 10.1109/TCBB.2022.3190265.
- [20] M. Umair dkk., "Detection of COVID-19 using transfer learning and Grad-CAM visualization on indigenously collected X-ray dataset," *Sensors*, vol. 21, no. 17, hal. 1–22, Sep. 2021, doi: 10.3390/S21175813.
- [21] T. Ozturk dkk., "Automated detection of COVID-19 cases using deep neural networks with X-ray images," *Comput. Biol. Med.*, vol. 121, hal. 1–11, Jun. 2020, doi: 10.1016/J.COMPBIOMED.2020.103792.
- [22] Y. Zhang dkk., "An interpretability optimization method for deep learning networks based on Grad-CAM," *IEEE Internet Things J.*, vol. 12, no. 4, hal. 3961–3970, Feb. 2025, doi: 10.1109/JIOT.2024.3485765.
- [23] R. Yulvina dkk., "Hybrid vision transformer and convolutional neural network for multi-class and multi-label classification of tuberculosis anomalies on chest X-ray," *Computers*, vol. 13, no. 12, hal. 1–29, Des. 2024, doi: 10.3390/COMPUTERS13120343.
- [24] R.A. Zeineldin dkk., "Explainable hybrid vision transformers and convolutional network for multimodal glioma segmentation in brain MRI," *Sci. Rep.*, vol. 14, hal. 1–14, Feb. 2024, doi: 10.1038/S41598-024-54186-7.

- [25] M.K. Islam dkk., "Enhancing lung abnormalities diagnosis using hybrid DCNN-ViT-GRU model with explainable AI," *Image Vis. Comput.*, vol. 142, hal. 1–18, Feb. 2024, doi: 10.1016/J.IMAVIS.2024.104918.