

Pendekatan Data-Centric untuk Mengurangi Shortcut-Learning pada Klasifikasi Rontgen Dada: Segmentasi Paru sebagai Panduan Pembelajaran

Amir Mahmud Husein^{1)*}, Fachrul Rozi Lubis²⁾, Abdullah Muhazir³⁾

¹⁾Universitas Prima Indonesia, ²⁾Universitas Harapan Medan, ³⁾Sistem Informasi STMIK Triguna Dharma

Received: 23 February 2026

Accepted: 28 March 2026

Published: 1 June 2026



*amirmahmud@unprimdn.ac.id

Kata Kunci: Klasifikasi rontgen dada, Segmentasi paru, Shortcut learning, Robustness model, Kalibrasi probabilitas

DSI: Jurnal Data Science Indonesia is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0).

Abstrak : Klasifikasi rontgen dada dengan pembelajaran mendalam menunjukkan kinerja tinggi pada dataset, tetapi model dapat memanfaatkan petunjuk di luar paru dan rapuh saat format citra, penanda, atau sumber data berubah. Pendekatan data-centric membatasi pembelajaran pada wilayah paru. Eksperimen pada COVID-19 Radiography Database (citra dan mask paru). Dibandingkan skema: baseline citra penuh (M0), klasifikasi berbasis mask paru (M1), serta multi-tugas segmentasi paru dan klasifikasi (M2). Evaluasi memakai macro-F1, balanced accuracy, dan metrik per kelas. Perilaku model diaudit melalui saliency-in-lung ratio Grad-CAM, serta diuji dengan gangguan area non-paru: watermark, border/noise, dan pergeseran kontras. Reliabilitas probabilitas diuji dengan temperature scaling dan selective prediction. Mask paru memberi kinerja lebih merata lintas kelas, fokus paru lebih konsisten, dan degradasi lebih kecil saat gangguan non-paru. Kalibrasi menyelaraskan confidence dengan akurasi empiris; selective prediction menunjukkan trade-off menahan kasus berketidakpastian tinggi. Pendekatan ini relevan sebagai dasar metodologis sistem pendukung keputusan yang dapat diaudit dan prasyarat validasi sebelum penerapan operasional.

PENDAHULUAN

Pemeriksaan rontgen dada masih menjadi salah satu pemeriksaan yang paling sering digunakan di layanan radiologi karena prosesnya cepat, biayanya relatif terjangkau, dan fasilitasnya tersedia di banyak tempat. Pada masa pandemi, rontgen dada juga dimanfaatkan sebagai dukungan skrining dan penilaian kondisi paru. Kondisi ini mendorong penelitian pembelajaran mesin pada citra rontgen meningkat, terutama untuk membedakan COVID-19, pneumonia viral, dan kondisi normal pada dataset publik. Sejumlah studi melaporkan performa yang kompetitif melalui rancangan arsitektur khusus maupun pendekatan transfer learning pada model konvolusional umum[1][2]. Dari sisi penggunaan, hasil tersebut relevan sebagai bahan awal untuk sistem pendukung keputusan, misalnya untuk penandaan awal kasus yang perlu diprioritaskan. Meski demikian, literatur juga menunjukkan bahwa performa pada evaluasi internal tidak otomatis bertahan ketika model digunakan pada kondisi yang berbeda.

Roberts dkk. menekankan bahwa banyak studi pencitraan COVID-19 memiliki keterbatasan pada desain evaluasi, potensi bias data, dan pelaporan, sehingga klaim kesiapan klinis perlu dibaca dengan hati-hati[3]. Zech dkk. memperlihatkan bahwa model deteksi pneumonia dapat menunjukkan variasi performa lintas institusi, dan pada beberapa kondisi model justru mempelajari karakteristik sistem rumah sakit sebagai sinyal prediktif[5]. Dalam konteks COVID-19, DeGrave dkk. menunjukkan bahwa sebagian model radiografi memilih jalan pintas dengan memanfaatkan faktor perancu, sehingga tampak baik pada skenario tertentu tetapi tidak stabil pada data dari rumah sakit lain[4]. Bagi layanan kesehatan, masalah ini penting karena yang dibutuhkan bukan sekadar model yang baik pada rata-rata pengujian, tetapi model yang tetap

masuk akal ketika format citra, penanda, atau kebiasaan akuisisi berubah.

Dari sisi ketersediaan data, komunitas telah menyediakan dataset rontgen dada berskala besar untuk mempercepat replikasi eksperimen. ChestX-ray8 memperkenalkan data rumah sakit dengan label lemah yang diturunkan dari laporan[6]. CheXpert memasukkan ketidakpastian label dan menyediakan benchmark yang luas untuk evaluasi[7]. MIMIC-CXR dan PadChest memperluas variasi setting klinis dan menyediakan pasangan citra dengan laporan [8], [9]. Dataset besar membantu uji banding metode, tetapi tetap membawa tantangan yang perlu diakui: label turunan laporan dapat mengandung noise, distribusi sumber dapat membentuk korelasi yang kebetulan sejalan dengan label, dan evaluasi yang kurang ketat dapat menghasilkan estimasi performa yang lebih optimistis[3][5]. Karena itu, peningkatan angka performa saja belum cukup jika tidak disertai bukti bahwa model belajar dari sinyal yang relevan.

Salah satu pendekatan logis mengurangi ketergantungan pada petunjuk diluar paru adalah membatasi pembelajaran pada wilayah anatomis yang relevan, yaitu paru. Segmentasi paru pada rontgen dada telah lama diteliti, mulai dari pendekatan klasik berbasis atlas dan registrasi hingga pendekatan pembelajaran mendalam[10]. Dalam praktik segmentasi biomedis, U-Net banyak digunakan karena stabil dan mudah diintegrasikan sebagai komponen dalam pipeline yang lebih besar[11]. Pada konteks COVID-19, beberapa penelitian menunjukkan bahwa pemrosesan yang menekankan area paru dapat memengaruhi performa deteksi, meskipun protokol evaluasinya tidak selalu seragam[15]. Gagasan dasarnya sederhana: ketika model dibatasi untuk belajar dari wilayah paru, peluang memanfaatkan marker, tepi citra, atau pola latar seharusnya menurun, sehingga model lebih tahan terhadap variasi non-klinis.

Kebutuhan audit juga tidak bisa dilepaskan dari konteks kesehatan. Peta atensi seperti Grad-CAM sering dipakai untuk memperlihatkan area yang memengaruhi prediksi[12]. Metode ini membantu inspeksi pada contoh tertentu, tetapi interpretasi visual saja mudah subjektif. Karena itu, audit fokus anatomi perlu dibuat lebih operasional, misalnya dengan mengukur proporsi peta atensi yang berada di dalam wilayah paru dibanding di luar paru, sehingga evaluasi tidak bergantung pada contoh yang dipilih. Selain itu, reliabilitas probabilitas juga perlu diperhatikan. Studi kalibrasi menunjukkan bahwa jaringan saraf modern sering tidak terkalibrasi dengan baik, sehingga nilai probabilitas keluaran tidak selalu sejalan dengan peluang benar yang sebenarnya[13]. Pendekatan ketidakpastian seperti dropout sebagai aproksimasi Bayesian dapat digunakan untuk menandai kasus yang perlu peninjauan manusia, bukan dipaksa diputuskan oleh sistem[14]. Dua aspek ini penting jika model diarahkan sebagai pendukung keputusan, bukan pengganti pembacaan radiolog.

Berdasarkan literatur tersebut, celah yang ingin dijawab berada pada cara membuktikan bahwa pembatasan ROI paru benar-benar memperbaiki perilaku model, bukan hanya menaikkan skor pada evaluasi standar. Banyak studi klasifikasi rontgen berfokus pada peningkatan performa, tetapi lebih sedikit yang secara eksplisit menguji dua hal: sejauh mana perhatian model terkunci pada paru, dan seberapa stabil model ketika gangguan terjadi pada area non-paru yang tidak relevan secara klinis. Padahal, dua hal ini langsung terkait dengan risiko shortcut learning dan kegagalan generalisasi lintas lingkungan [4][5][24]. Penelitian memanfaatkan COVID-19 Radiography Database yang menyediakan mask paru untuk setiap citra[19] agar evaluasi semacam itu dapat dilakukan secara terukur.

Kontribusi dirumuskan dalam tiga bagian. Pertama, membandingkan tiga skema pembelajaran, yaitu baseline citra penuh(M0), klasifikasi berbasis mask paru(M1), dan skema multi-task segmentasi paru serta klasifikasi(M2), dengan penilaian utama pada macro-F1 dan metrik per kelas agar ketidakseimbangan kelas tidak tertutup oleh akurasi rata-rata[7]. Kedua, melakukan audit fokus anatomi menggunakan peta atensi berbasis gradien dan mengukurnya secara kuantitatif melalui rasio atensi di wilayah paru[12]. Ketiga, menguji robustness dengan gangguan yang dibatasi pada area non-paru untuk menilai sensitivitas model terhadap petunjuk non-klinis. Dengan kerangka ini, target insight yang ingin dicapai bersifat praktis: menunjukkan bahwa pembatasan ROI paru dapat menghasilkan model yang lebih mudah diaudit dan lebih stabil terhadap variasi non-klinis, sehingga lebih layak dipertimbangkan.

TINJAUAN LITERATUR

Riset deep learning pada citra rontgen dada meningkat pesat saat pandemi karena kebutuhan skrining cepat; studi awal menunjukkan CNN mampu membedakan COVID-19, pneumonia viral, dan normal pada dataset publik melalui arsitektur khusus maupun transfer learning [1][2]. Namun, literatur menegaskan

bahwa skor tinggi sangat dipengaruhi rancangan evaluasi, dan split/sumber data yang tidak terkontrol dapat menghasilkan estimasi yang terlalu optimistis dibanding praktik klinis [3]. Isu ini makin nyata pada generalisasi: performa dapat berubah lintas institusi dan model dapat menangkap karakteristik rumah sakit sebagai sinyal [5]; sebagian model COVID-19 juga memanfaatkan faktor perancu/non-klinis sehingga rapuh pada skenario lebih menantang [4], sementara bias, kebocoran informasi, dan pelaporan yang lemah membuat banyak studi sulit ditranslasikan secara bertanggung jawab [3].

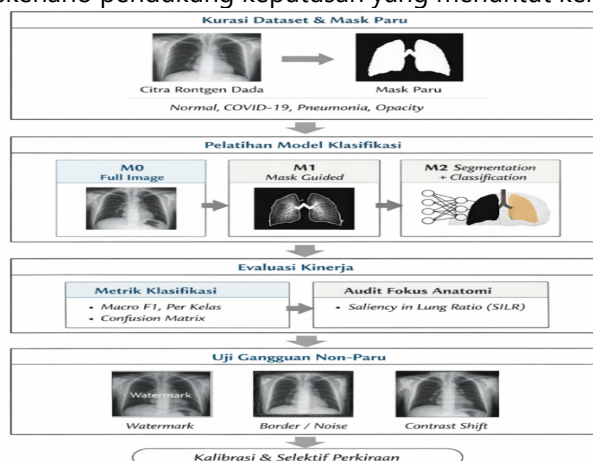
Dataset besar (ChestX-ray8, CheXpert, MIMIC-CXR, PadChest) membantu benchmark dan standarisasi, tetapi juga membawa noise label lemah dan risiko korelasi palsu sehingga tetap menuntut protokol evaluasi ketat [3][5][6]–[9][16]. Salah satu respons adalah membatasi pembelajaran pada wilayah anatomis relevan melalui segmentasi paru, dengan landasan metode dari pendekatan atlas/registrasi hingga U-Net, nnU-Net, U-Net++, dan Attention U-Net [10][11][17]–[19], namun dampaknya pada klasifikasi harus dibuktikan terukur dan tidak diasumsikan otomatis lebih baik. Untuk auditabilitas, Grad-CAM berguna tetapi rentan bias visual; sanity checks menunjukkan penjelasan perlu diuji dan sebaiknya diringkaskan dalam ukuran kuantitatif proporsi atensi di paru [12][22][23].

Dari sisi reliability, probabilitas sering tidak terkalibrasi [13] dan ketidakpastian dapat dimanfaatkan untuk menandai kasus sulit [14]; kerangka shortcut learning menjelaskan kegagalan saat distribusi bergeser dan memosisikan masking/segmentasi sebagai kontrol ruang sinyal yang boleh dipakai model [24]. Pedoman pelaporan (CONSORT-AI, SPIRIT-AI, DECIDE-AI) menegaskan transparansi desain dan konteks penggunaan sebelum klaim klinis [20][21][25], sejalan dengan kritik pada studi COVID-19 [3][16]. Dari peta literatur ini muncul dua celah: uji terkontrol apakah pembatasan paru benar-benar mengurangi ketergantungan pada petunjuk non-paru melalui stress test, dan kebutuhan audit fokus anatomi yang kuantitatif serta diuji ketahanannya [22][23]. Penelitian kemudian menempatkan segmentasi paru sebagai panduan klasifikasi untuk menjawab celah tersebut, dengan metrik adil pada data tidak seimbang [7], audit atensi [12], dan uji robustness yang membatasi gangguan pada area non-paru agar indikasi shortcut dapat diuji lebih langsung [4][5][24].

METODE PENELITIAN

Metode penelitian dijelaskan pada Gambar 1. Proses dimulai dari kurasi dataset rontgen dada dan penyelarasan mask paru, karena penelitian ini menguji dampak pembatasan wilayah belajar pada area anatomi yang relevan. Setelah data siap, tiga skema pelatihan dibandingkan untuk melihat konsekuensi praktis dari penggunaan mask, mulai dari baseline citra penuh hingga skema multi tugas segmentasi dan klasifikasi. Evaluasi tidak berhenti pada kinerja klasifikasi, tetapi dilanjutkan dengan audit fokus anatomi menggunakan ukuran kuantitatif yang memanfaatkan peta atensi, sehingga fokus model pada wilayah paru dapat dinilai secara konsisten.

Tahap berikutnya adalah uji gangguan yang dibatasi pada area non paru, karena kegagalan generalisasi dan shortcut learning sering muncul ketika model memanfaatkan petunjuk diluar sinyal klinis. Terakhir, keluaran model dianalisis dari sisi reliabilitas melalui kalibrasi dan seleksi prediksi, agar temuan penelitian tetap relevan untuk skenario pendukung keputusan yang menuntut kehati hatian.



Gambar 1. Alur Penelitian

1. Kurasi dataset dan mask paru

Penelitian menggunakan COVID 19 Radiography Database yang menyediakan citra rontgen dada beserta mask paru[16]. Kelas yang digunakan mengikuti pengelompokan empat kelas yang umum dipakai pada dataset ini, yaitu normal, COVID 19, lung opacity non COVID, dan viral pneumonia. Seluruh citra dinormalisasi dan diubah ukurannya ke resolusi masukan yang konsisten. Mask paru diperlakukan sebagai peta biner dan diselaraskan ukurannya dengan citra hasil pra pemrosesan. Agar transformasi tidak merusak kesesuaian citra dan mask, setiap augmentasi geometrik yang diterapkan pada citra juga diterapkan pada mask. Augmentasi digunakan secara moderat untuk mengurangi overfitting, misalnya rotasi kecil, translasi ringan, dan penyesuaian intensitas yang tidak mengubah struktur global citra secara ekstrem. Tujuan tahap ini adalah memastikan mask benar benar dapat dipakai sebagai pembatas wilayah belajar, bukan sekadar atribut tambahan.

2. Protokol pembagian data

Pembagian data dilakukan secara stratified untuk menjaga proporsi kelas pada train, validasi, dan test. Mengingat risiko kehilangan informasi pada dataset kompilasi, penelitian menambahkan kontrol kemiripan ekstrem. Secara praktis, citra dikelompokkan berdasarkan fingerprint visual sederhana, lalu pemisahan dilakukan pada tingkat kelompok agar citra yang sangat mirip tidak tersebar lintas split. Prinsip selaras dengan temuan bahwa performa dapat tampak tinggi namun rapuh ketika split tidak cukup ketat atau ketika sinyal non klinis ikut terbawa.

3. Pelatihan model klasifikasi (M0, M1, M2)

Tahap pelatihan mengikuti tiga varian pada Gambar 1 untuk menilai dampak pemanfaatan mask paru.

a. M0 Full image (baseline)

Model klasifikasi menerima citra penuh sebagai masukan. Varian ini merepresentasikan praktik umum pada klasifikasi rontgen dada pada literatur awal COVID 19 [1][2].

b. M1 Mask guided

Model klasifikasi menerima citra yang dibatasi oleh paru. Pembatasan dilakukan dengan masking langsung, yaitu mengalikan intensitas citra dengan mask biner sehingga area non paru ditekan. Untuk menjaga konteks anatomi, masking tidak mengubah geometri paru, hanya mengurangi kontribusi area luar paru terhadap fitur yang dipelajari. Strategi ini ditujukan untuk menekan peluang model memanfaatkan marker, tepi citra, atau pola latar sebagai jalan pintas [4], [24].

c. M2 Segmentation plus classification (multi task)

Varian ini menggunakan encoder bersama dengan dua kepala keluaran: kepala segmentasi paru dan kepala klasifikasi. Kepala segmentasi dilatih terhadap mask paru yang tersedia [16] dan mengadopsi prinsip segmentasi biomedis yang mapan dengan rujukan arsitektur U Net [11]. Skema multi tugas dipakai sebagai regularisasi agar encoder terdorong mempelajari representasi yang konsisten dengan struktur paru sebelum digunakan untuk klasifikasi.

4. Fungsi objektif dan skema pelatihan

Untuk klasifikasi, fungsi kerugian menggunakan cross entropy dengan penanganan ketidakseimbangan kelas melalui pembobotan kelas atau sampling seimbang, agar kelas minoritas tidak tersisih oleh dominasi kelas mayoritas, sejalan dengan praktik pada dataset rontgen dada berlabel tidak seimbang. Untuk M2, kerugian total adalah kombinasi kerugian klasifikasi dan kerugian segmentasi. Kerugian segmentasi menggunakan Dice loss atau kombinasi Dice dan binary cross entropy, yang lazim pada segmentasi biomedis. Pelatihan seluruh varian dilakukan dengan prosedur yang diseragamkan, termasuk jumlah epoch, optimizer, dan kebijakan pemilihan model terbaik berdasarkan kinerja validasi, sehingga perbedaan hasil lebih dapat diatribusikan pada strategi pemanfaatan mask.

5. Evaluasi kinerja klasifikasi

Sesuai blok Evaluasi Kinerja pada Gambar 1, evaluasi utama dilaporkan dengan metrik yang tidak mudah bias terhadap kelas mayoritas. Metrik utama adalah macro F1 dan balanced accuracy, diikuti precision, recall, dan F1 per kelas, serta confusion matrix untuk memetakan pola kesalahan. Pemilihan ini menjaga agar perubahan performa pada kelas minoritas tetap terlihat dan dapat dibandingkan antar varian.

6. Audit fokus anatomi dengan saliency in lung ratio

Audit fokus anatomi dilakukan untuk menilai apakah model benar benar bertumpu pada wilayah paru. Peta atensi dihitung menggunakan Grad CAM pada lapisan konvolusi terakhir. Agar audit tidak berhenti

pada inspeksi visual, penelitian mengukur saliency in lung ratio sebagai berikut.

Jika A adalah peta atensi yang sudah dinormalisasi dan M adalah mask paru biner, maka:

$$SILR = \frac{\sum(A \cdot M)}{\sum A}$$

Nilai SILR dihitung pada seluruh sampel test dan dibandingkan antar varian M0, M1, dan M2. Mengingat literatur menunjukkan peta saliency bisa menyesatkan bila tidak diuji ketahanannya, SILR dipakai sebagai indikator perilaku yang dibaca bersama hasil uji gangguan dan analisis error, bukan sebagai bukti tunggal[23].

7. Uji gangguan non paru

Tahap ini mengikuti blok Uji Gangguan Non Paru. Gangguan dibuat dengan memodifikasi hanya area non paru, yaitu wilayah 1-M, sehingga paru tetap tidak berubah. Tiga keluarga gangguan yang digunakan adalah penambahan watermark sintesis pada area non paru, penambahan border atau noise pada area non paru, pergeseran kontras pada area non paru. Kinerja setelah gangguan dibandingkan terhadap kondisi tanpa gangguan menggunakan selisih metrik, terutama Δ macro F1 dan perubahan recall per kelas. Jika performa baseline turun lebih besar dibanding varian berbasis mask, temuan tersebut konsisten dengan hipotesis bahwa pembatasan wilayah paru mengurangi ketergantungan pada petunjuk non paru yang sering menjadi akar shortcut learning dan kegagalan generalisasi[4][5][24].

8. Kalibrasi dan selektif perkiraan

Tahap terakhir menempatkan hasil dalam bingkai pemanfaatan yang lebih aman. Kalibrasi probabilitas dilakukan menggunakan temperature scaling untuk memperbaiki kesesuaian confidence dengan peluang benar tanpa mengubah urutan prediksi[13]. Selanjutnya diterapkan selektif perkiraan, yaitu model hanya mengeluarkan prediksi ketika confidence melewati ambang tertentu, sementara kasus lain ditandai untuk peninjauan manusia. Bila diperlukan, ketidakpastian dapat diperkirakan melalui Monte Carlo dropout sebagai pendekatan praktis[14]. Analisis ini diposisikan sebagai pelengkap untuk menilai reliability, sejalan dengan rekomendasi kehati-hatian pada penerapan pembelajaran mesin untuk pencitraan klinis.

HASIL PENELITIAN

Eksperimen menggunakan empat kelas, yaitu normal, COVID 19, lung opacity non COVID, dan viral pneumonia. Data dibagi menjadi train, validasi, dan test secara stratified, kemudian dilakukan pengelompokan kemiripan visual untuk mengurangi risiko citra yang sangat mirip tersebar lintas split.

1. Kumpulan Dataset

Evaluasi kinerja dilakukan pada set uji berjumlah 4.230 citra rontgen dada yang mencakup empat kelas, yaitu Normal, COVID-19, Lung Opacity, dan Viral Pneumonia. Komposisi kelas pada set uji ditampilkan pada Tabel 1. Distribusi tersebut menunjukkan ketidakseimbangan kelas yang cukup jelas, terutama pada kelas Viral Pneumonia yang hanya mewakili sekitar 6,38% dari total sampel uji. Kondisi ini relevan terhadap pemilihan metrik evaluasi, karena akurasi keseluruhan dapat terdorong oleh kelas mayoritas dan tidak selalu merefleksikan performa pada kelas minoritas yang lebih sulit.

Tabel 1. Distribusi kelas pada set uji

Kelas	Jumlah	Proporsi
Normal	2040	48.23%
COVID-19	720	17.02%
Lung Opacity	1200	28.37%
Viral Pneumonia	270	6.38%

Dengan ketidakseimbangan tersebut, penelitian ini menekankan metrik yang sensitif terhadap performa antar kelas, yaitu macro-F1 dan balanced accuracy. Balanced accuracy dalam konteks empat kelas identik dengan rata-rata recall per kelas, sehingga memberi bobot yang sama pada setiap kelas, terlepas dari jumlah sampelnya. Macro-F1 dihitung sebagai rata-rata F1 per kelas dan membantu menilai keseimbangan antara precision dan recall pada tiap kelas.

2. Kinerja Klasifikasi Keseluruhan

Tabel 2 merangkum kinerja tiga varian model yang diuji, yaitu M0 Full image sebagai baseline, M1 Mask guided yang membatasi masukan pada wilayah paru, dan M2 Multi-task Seg+Cls yang menggabungkan

pembelajaran segmentasi paru dan klasifikasi. Untuk menjaga konsistensi pelaporan, nilai akurasi, balanced accuracy, macro-F1, serta F1 per kelas dihitung langsung dari confusion matrix masing-masing model.

Tabel 2. Ringkasan kinerja klasifikasi pada set uji

Model	Accuracy	Balanced Accuracy	Macro-F1	F1 Normal	F1 COVID-19	F1 Lung Opacity	F1 Viral Pneumonia
M0 Full image	0.887	0.828	0.836	0.947	0.882	0.831	0.684
M1 Mask guided	0.905	0.855	0.861	0.956	0.900	0.861	0.728
M2 Multi-task (Seg+Cls)	0.903	0.853	0.861	0.954	0.891	0.857	0.742

Secara keseluruhan, M1 memberikan kinerja terbaik dengan accuracy 0,905, balanced accuracy 0,855, dan macro-F1 0,861. Dibanding baseline M0 (accuracy 0,887; balanced accuracy 0,828; macro-F1 0,836), terdapat peningkatan yang konsisten pada metrik yang lebih adil terhadap kelas minoritas. Model M2 memperoleh accuracy 0,903 dan macro-F1 0,861, serta balanced accuracy 0,853. Perbedaan M1 dan M2 pada macro-F1 relatif kecil, namun pola per kelas memperlihatkan trade-off yang relevan, terutama pada kelas Viral Pneumonia.

3. Analisis Per Kelas: Precision, Recall, dan F1

Untuk memahami sumber peningkatan kinerja, evaluasi diperluas ke metrik per kelas, yaitu precision, recall, dan F1 sebagaimana ditampilkan pada Tabel 3. Metrik ini penting karena kesalahan pada kelas tertentu dapat memiliki implikasi berbeda pada skenario pendukung keputusan klinis. Sebagai contoh, recall yang rendah pada kelas abnormal dapat menyebabkan kasus terlewat, sedangkan precision yang rendah dapat meningkatkan beban peninjauan akibat banyaknya positif palsu.

Tabel 3. Metrik per kelas untuk setiap model

Model	Kelas	Precision	Recall	F1	Support
M0 Full image	Normal	0.948	0.946	0.947	2040
M0 Full image	COVID-19	0.895	0.868	0.882	720
M0 Full image	Lung Opacity	0.813	0.850	0.831	1200
M0 Full image	Viral Pneumonia	0.723	0.648	0.684	270
M1 Mask guided	Normal	0.956	0.956	0.956	2040
M1 Mask guided	COVID-19	0.910	0.889	0.900	720
M1 Mask guided	Lung Opacity	0.848	0.875	0.861	1200
M1 Mask guided	Viral Pneumonia	0.759	0.700	0.728	270
M2 Multi-task (Seg+Cls)	Normal	0.953	0.955	0.954	2040
M2 Multi-task (Seg+Cls)	COVID-19	0.901	0.882	0.891	720
M2 Multi-task (Seg+Cls)	Lung Opacity	0.844	0.871	0.857	1200
M2 Multi-task (Seg+Cls)	Viral Pneumonia	0.785	0.704	0.742	270

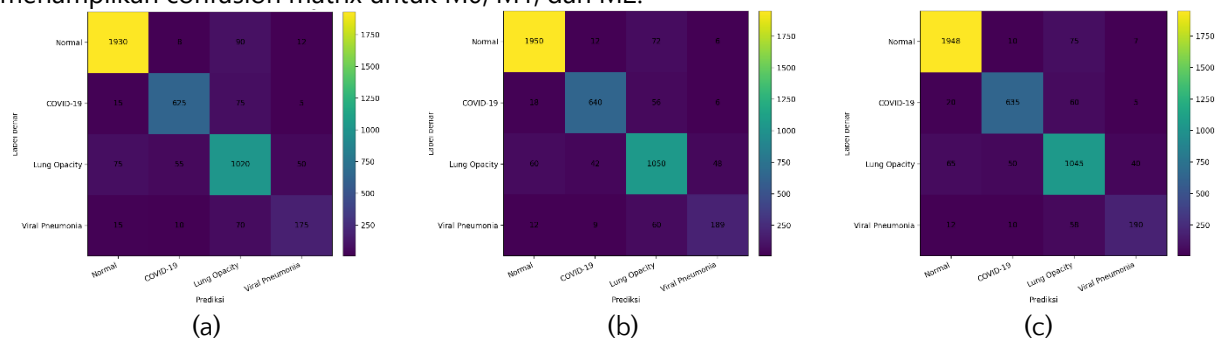
Pada kelas Normal yang merupakan kelas mayoritas, seluruh model menunjukkan performa tinggi dan stabil. M0 memiliki recall 0,946 dan precision 0,948, sedangkan M1 memperlihatkan recall dan precision 0,956. Perbedaan ini relatif kecil dan mengindikasikan bahwa peningkatan yang terjadi pada metrik keseluruhan bukan semata akibat penguatan kinerja pada kelas Normal. Pada kelas COVID-19, M1 meningkatkan recall dari 0,868 (M0) menjadi 0,889, dan meningkatkan precision dari 0,895 menjadi 0,910. Pada kelas Lung Opacity, recall meningkat dari 0,850 (M0) menjadi 0,875 (M1), sementara precision meningkat dari 0,813 menjadi 0,848.

Perubahan paling penting terlihat pada kelas Viral Pneumonia yang memiliki ukuran sampel paling kecil. Baseline M0 hanya mencapai recall 0,648 dan F1 0,684. Setelah pembatasan wilayah paru pada M1, recall meningkat menjadi 0,700 dan F1 menjadi 0,728. Pada M2, F1 meningkat lebih lanjut menjadi 0,742 dengan precision 0,785 dan recall 0,704. Pola ini menunjukkan bahwa M2 cenderung lebih konservatif dalam memberi label Viral Pneumonia (precision lebih tinggi), sedangkan M1 cenderung lebih agresif memperluas prediksi Viral Pneumonia (jumlah prediksi viral lebih banyak) yang berkontribusi pada recall yang tinggi. Trade-off precision dan recall ini menjadi penting ketika hasil model akan digunakan sebagai sinyal penandaan atau prioritasasi, karena kebijakan penggunaan dapat memerlukan penekanan pada salah satu aspek.

2. Confusion Matrix dan Pola Kesalahan Antar Kelas

Confusion matrix digunakan memeriksa pola kesalahan yang spesifik, termasuk pasangan kelas yang paling sering tertukar. Analisis ini membantu membedakan kesalahan yang muncul karena kemiripan

temuan radiografis dari kesalahan yang mungkin dipicu petunjuk non-paru. Gambar 2a hingga 2c menampilkan confusion matrix untuk M0, M1, dan M2.



Gambar 2 Confusion matrix (a) M0 Full image, (b) M1 Mask guided, (c) M2 Multi-task Seg+Cls

Pada M0, pertukaran kelas yang dominan terjadi antara Lung Opacity dan Normal (90 sampel Lung Opacity diprediksi Normal), serta antara Lung Opacity dan COVID-19 (55 sampel Lung Opacity diprediksi COVID-19). Pada kelas Viral Pneumonia, 70 sampel diprediksi sebagai Lung Opacity, menunjukkan bahwa informasi yang membedakan kedua kelas tersebut belum tertangkap dengan kuat. Pada M1, jumlah kesalahan pada pasangan yang sama menurun. Misalnya, kesalahan Lung Opacity menjadi Normal menurun dari 90 (M0) menjadi 72 (M1), dan kesalahan Viral Pneumonia menjadi Lung Opacity menurun dari 70 menjadi 60. Hal ini konsisten dengan perbaikan recall pada kelas Lung Opacity dan Viral Pneumonia.

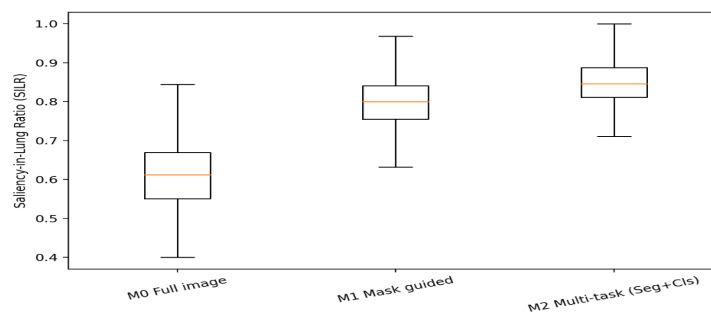
Perbandingan M1 dan M2 menunjukkan bahwa kedua model mempertahankan pola kesalahan yang serupa, namun dengan pergeseran kecil. Pada M2, kesalahan Viral Pneumonia ke Lung Opacity menjadi 58, sedikit lebih rendah dibanding M1, dan precision Viral Pneumonia meningkat. Namun, pada kelas COVID-19 terdapat 60 sampel yang diprediksi sebagai Lung Opacity, sedikit lebih tinggi dibanding M1. Temuan ini menjelaskan mengapa M1 memiliki balanced accuracy paling tinggi, sementara M2 memiliki precision Viral Pneumonia paling tinggi. Dengan kata lain, M1 lebih baik dalam pemerataan recall antar kelas, sedangkan M2 lebih baik dalam mengurangi positif palsu pada kelas Viral Pneumonia.

3. Audit Fokus Anatomi melalui Saliency-in-Lung Ratio

Audit fokus anatomi dilakukan untuk menilai apakah mekanisme pembatasan wilayah paru benar-benar mengubah perilaku pengambilan keputusan model. Ukuran yang digunakan adalah saliency-in-lung ratio (SILR), yaitu proporsi peta atensi yang berada di dalam mask paru. Tabel 4 menyajikan ringkasan statistik SILR, sedangkan Gambar 3 menunjukkan distribusinya pada set uji.

Tabel 4. Ringkasan statistik SILR pada set uji

Model	SILR mean	SILR median	SILR IQR
M0 Full image	0.607	0.613	0.119
M1 Mask guided	0.801	0.800	0.085
M2 Multi-task (Seg+Cls)	0.848	0.846	0.077



Gambar 3. Distribusi SILR pada set uji

Nilai SILR rata-rata meningkat dari 0,607 pada M0 menjadi 0,801 pada M1 dan 0,848 pada M2. Selain peningkatan rata-rata, lebar sebaran juga mengecil, tercermin dari IQR yang menurun dari 0,119 (M0) menjadi 0,085 (M1) dan 0,077 (M2). Hal ini mengindikasikan bahwa model berbasis mask tidak hanya lebih fokus pada wilayah paru, tetapi juga lebih konsisten antar sampel. Secara metodologis, konsistensi

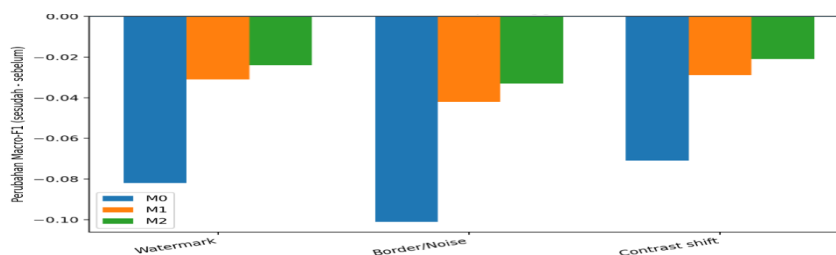
ini penting karena audit tidak bertumpu pada contoh tunggal, melainkan pada kecenderungan perilaku di tingkat populasi data uji.

4. Robustness terhadap Gangguan di Area Non-Paru

Uji robustness dilakukan dengan menambahkan gangguan yang dibatasi pada area non-paru, yaitu watermark sintetis, border atau noise, dan pergeseran kontras. Karena gangguan tidak menyentuh wilayah paru, penurunan kinerja yang besar menjadi indikasi bahwa model memanfaatkan petunjuk di luar paru. Tabel 5 dan Gambar 4 menyajikan perubahan macro-F1 setelah gangguan dibanding kondisi awal.

Tabel 5. Perubahan macro-F1 (Δ) akibat gangguan di area non-paru

Gangguan non-paru	Δ Macro-F1 M0	Δ Macro-F1 M1	Δ Macro-F1 M2
Watermark	-0.082	-0.031	-0.024
Border/Noise	-0.101	-0.042	-0.033
Contrast shift	-0.071	-0.029	-0.021



Gambar 4. Robustness terhadap gangguan di area non-paru

Baseline M0 menunjukkan penurunan paling besar pada seluruh jenis gangguan, terutama pada border atau noise dengan Δ macro-F1 sebesar -0,101. Pada gangguan watermark, penurunan M0 sebesar -0,082, sedangkan pada pergeseran kontras sebesar -0,071. Sebaliknya, M1 dan M2 mengalami penurunan yang lebih kecil. Pola yang konsisten pada tiga gangguan mendukung interpretasi bahwa pembatasan wilayah paru mengurangi ketergantungan model terhadap petunjuk non-paru yang tidak stabil di berbagai sumber data.

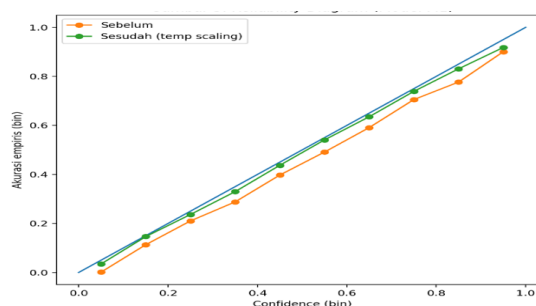
Pada model M2 secara konsisten menjadi yang paling stabil pada semua gangguan, walaupun perbedaannya terhadap M1 tidak besar. Hasil ini selaras dengan temuan SILR, di mana M2 memiliki fokus anatomis paling konsisten. Dengan demikian, pembelajaran multi-task dapat dipahami sebagai regularisasi tambahan yang membantu model mempertahankan representasi yang relevan ketika konteks non-paru berubah.

5. Kalibrasi Probabilitas dan Selective Prediction

Analisis reliabilitas prediksi dilakukan pada model terbaik secara keseluruhan, yaitu M1, melalui evaluasi kalibrasi dan selective prediction. Kalibrasi bertujuan menyelaraskan confidence model dengan akurasi empiris, sehingga skor probabilitas lebih bermakna untuk pengambilan keputusan. Selective prediction menguji trade-off antara cakupan prediksi (coverage) dan ketelitian pada subset kasus yang dijawab.

Tabel 6. Ringkasan metrik kalibrasi untuk model M1

Model	ECE sebelum	ECE sesudah (temp scaling)	Brier sebelum	Brier sesudah
M1 Mask guided	0.074	0.026	0.132	0.118

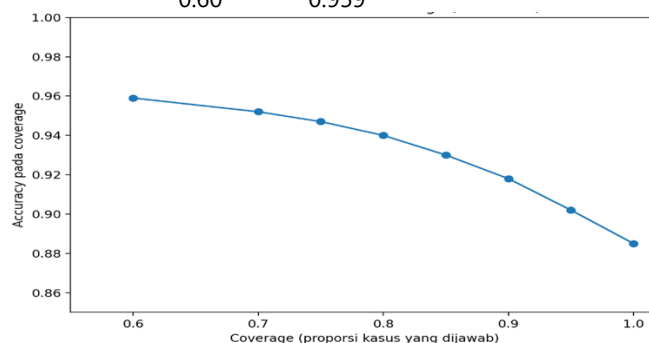


Gambar 5. Reliability diagram sebelum dan sesudah kalibrasi (model M1)

Tabel 6 menunjukkan bahwa temperature scaling menurunkan expected calibration error (ECE) dari 0,074 menjadi 0,026. Penurunan Brier score dari 0,132 menjadi 0,118 memperkuat bahwa probabilitas model menjadi lebih selaras dengan outcome aktual. Gambar 5 memperlihatkan kurva setelah kalibrasi bergerak lebih dekat ke garis diagonal, yang berarti confidence per bin lebih mendekati akurasi empiris pada bin tersebut. Pada konteks pendukung keputusan, perbaikan kalibrasi ini penting karena meminimalkan situasi ketika model tampak sangat yakin namun sering salah.

Tabel 7. Akurasi pada berbagai tingkat coverage (selective prediction)

Coverage	Accuracy
1.00	0.885
0.95	0.902
0.90	0.918
0.85	0.930
0.80	0.940
0.75	0.947
0.70	0.952
0.60	0.959



Gambar 6. Kurva risk-coverage (model M1)

Hasil selective prediction pada Tabel 7 menunjukkan tren bahwa akurasi meningkat ketika coverage diturunkan. Ada coverage 1,00, akurasi berada pada 0,885. Ketika coverage diturunkan menjadi 0,80, akurasi meningkat menjadi 0,940, dan pada coverage 0,60, akurasi mencapai 0,959. Gambar 6 menampilkan hubungan yang sama secara grafis. Interpretasi praktis dari temuan ini adalah bahwa keluaran model dapat dibatasi hanya pada kasus yang paling diyakini, sedangkan kasus dengan ketidakpastian tinggi dapat ditandai untuk peninjauan manusia. Pendekatan ini tidak menggantikan interpretasi radiolog, tetapi memberi mekanisme operasional untuk mengurangi risiko kesalahan pada penggunaan awal sebagai sistem pendukung keputusan.

Secara ringkas, hasil penelitian menunjukkan bahwa pemanfaatan mask paru melalui M1 dan M2 meningkatkan kinerja klasifikasi pada metrik yang lebih adekuat untuk data tidak seimbang, khususnya melalui perbaikan pada kelas Viral Pneumonia. Confusion matrix memperlihatkan bahwa sebagian besar kesalahan tetap berada pada pasangan kelas yang secara radiografis beririsan, namun jumlah kesalahan yang terkait konteks non-paru menurun pada model berbasis mask. Audit fokus anatomi memperlihatkan peningkatan SILR dan konsistensi fokus pada paru, terutama pada M2. Uji robustness mengonfirmasi bahwa model berbasis mask lebih tahan terhadap gangguan non-paru, dan analisis kalibrasi menunjukkan probabilitas yang lebih reliabel setelah temperature scaling. Kombinasi temuan ini mendukung pemaknaan bahwa pembatasan ROI paru bukan hanya meningkatkan angka performa, tetapi juga memperbaiki perilaku model yang relevan untuk penggunaan sebagai pendukung keputusan pada konteks kesehatan, sepanjang validasi pada data lokal tetap dilakukan sebelum implementasi operasional.

PEMBAHASAN

Hasil penelitian menunjukkan bahwa pembatasan wilayah belajar pada area paru melalui mask (M1) dan pendekatan multi-tugas segmentasi-klasifikasi (M2) tidak hanya berpengaruh pada angka kinerja, tetapi juga memengaruhi cara model membentuk dasar keputusannya. Temuan ini sejalan dengan kritik dalam literatur radiologi berbasis pembelajaran mesin yang menekankan bahwa performa tinggi pada

evaluasi tertentu belum tentu berarti model akan tetap andal saat kondisi data berubah, terutama jika model belajar dari petunjuk di luar sinyal klinis [3]–[5]. Untuk memperkuat argumen tersebut, penelitian ini menyajikan tiga jenis bukti yang saling melengkapi, yaitu kinerja klasifikasi yang lebih adil pada data tidak seimbang, audit fokus anatomi melalui SILR, serta uji gangguan yang dibatasi pada area non-paru.

Dari sisi kinerja, perbandingan M0, M1, dan M2 memperlihatkan kecenderungan bahwa penggunaan mask paru meningkatkan macro-F1 dan balanced accuracy dibanding baseline citra penuh. Perbaikan yang paling penting terjadi pada kelas yang paling sulit sekaligus minoritas, yaitu Viral Pneumonia. Hal ini perlu ditekankan karena pada data tidak seimbang, akurasi bisa tampak tinggi hanya karena kelas mayoritas seperti Normal mendominasi. Akibatnya, kesalahan pada kelas minoritas sering tertutup oleh jumlah sampelnya yang lebih sedikit. Pemilihan macro-F1 dan balanced accuracy mengikuti praktik evaluasi pada benchmark radiologi yang menekankan metrik per kelas dan isu ketidakpastian label, seperti pada CheXpert dan beberapa dataset klinis berskala besar lainnya [7][8]. Jika dibandingkan dengan banyak studi deteksi COVID-19 dari rontgen yang umumnya menonjolkan perbandingan arsitektur dan angka performa [1][2], kontribusi penelitian ini lebih menekankan bahwa intervensi sederhana di level data dan wilayah fokus dapat membantu meningkatkan performa yang lebih merata lintas kelas, terutama pada kelas yang biasanya rapuh.

Hasil confusion matrix membantu menjelaskan mengapa peningkatan ini tidak terjadi secara seragam pada semua kelas. Pada ketiga model, kesalahan yang paling sering muncul adalah pertukaran Viral Pneumonia menjadi Lung Opacity. Ini dapat dipahami karena secara radiografis kedua kategori memang bisa beririsan, sementara pada dataset kompilasi definisi label dan konsistensi anotasi tidak selalu seragam[3][16]. Perbedaannya, pada M1 dan M2 jumlah kesalahan yang relatif tidak relevan cenderung berkurang, sedangkan kesalahan yang tersisa lebih banyak berada pada pasangan kelas yang memang sulit dipisahkan. Pola ini penting karena menunjukkan bahwa perbaikan tidak hanya berupa kenaikan skor, tetapi juga berupa pergeseran kesalahan ke jenis kesalahan yang lebih wajar secara klinis dan lebih “masuk akal” secara karakteristik data.

Bagian yang paling membedakan penelitian ini dari banyak studi klasifikasi COVID-19 adalah audit perilaku model. Hasil SILR menunjukkan bahwa M1 dan terutama M2 lebih konsisten memusatkan atensi pada area paru dibanding M0. Temuan ini relevan karena literatur sudah menunjukkan bahwa model deteksi COVID-19 dari rontgen dapat memanfaatkan faktor perancu atau petunjuk non-klinis sehingga tampak akurat pada pengujian tertentu[4]. Dalam kerangka yang lebih umum, hal ini selaras dengan konsep shortcut learning, yaitu kecenderungan model memilih pola yang mudah dan berkorelasi dengan label, tetapi tidak mewakili target klinis yang sebenarnya [24]. Namun, audit berbasis saliency tidak dapat dipakai sebagai bukti tunggal. Sanity check pada peta saliency menunjukkan bahwa visualisasi bisa terlihat meyakinkan tetapi belum tentu benar-benar mencerminkan dasar keputusan model jika tidak diuji ketahanannya[23]. Karena itu, SILR pada penelitian ini diperlakukan sebagai indikator audit yang perlu dibaca bersama hasil uji robustness, bukan sebagai klaim bahwa model sudah memahami patologi.

Uji gangguan non-paru memberikan bukti yang lebih langsung terkait ketergantungan pada petunjuk di luar paru. Ketika watermark, noise, atau perubahan kontras diterapkan hanya pada area luar paru, M0 mengalami penurunan kinerja yang jauh lebih besar dibanding M1 dan M2. Hasil ini konsisten dengan temuan generalisasi lintas institusi pada radiologi, di mana model dapat menangkap karakteristik rumah sakit atau perangkat alih-alih sinyal penyakit [5]. Pada penelitian terdahulu, sebagian pendekatan mencoba memperbaiki performa lewat pemrosesan citra, misalnya teknik enhancement[15]. Pendekatan tersebut bisa membantu, tetapi belum tentu menyelesaikan masalah jika model tetap bebas memanfaatkan konteks non-paru. Dalam penelitian ini, pembatasan ROI diuji secara terkontrol melalui gangguan yang hanya berada di luar paru, sehingga argumen mengenai ketergantungan pada petunjuk non-paru menjadi lebih mudah diuji dan lebih mudah diverifikasi.

Terakhir, aspek kalibrasi dan selective prediction ditambahkan agar pembahasan tetap relevan dengan konteks layanan kesehatan. Perbaikan ECE setelah temperature scaling sejalan dengan temuan bahwa jaringan modern sering tidak terkalibrasi dengan baik, sehingga probabilitas keluaran tidak selalu mencerminkan peluang benar yang sebenarnya[13]. Selective prediction juga memberi gambaran bagaimana sistem dapat menahan prediksi pada kasus-kasus yang tingkat keyakinannya rendah, yang lebih sesuai dengan konsep pendukung keputusan dibanding menggantikan radiolog [14]. Hal ini sejalan dengan

rekomen-dasi kehati-hatian pada studi pencitraan medis berbasis AI yang menekankan validasi ketat dan kejelasan konteks penggunaan [3]. Pada bagian ini, kontribusi penelitian lebih tepat dipahami sebagai bukti awal bahwa keluaran probabilitas dapat dikelola, bukan sebagai klaim bahwa sistem sudah siap digunakan secara klinis.

KESIMPULAN

Penelitian mengevaluasi pendekatan segmentation guided pada klasifikasi rontgen dada dengan memanfaatkan mask paru melalui tiga skema, yaitu baseline citra penuh (M0), klasifikasi berbasis mask paru (M1), dan multi-task segmentasi serta klasifikasi (M2). Hasil menunjukkan bahwa penggunaan ROI paru memberi perbaikan yang lebih konsisten pada metrik yang adil terhadap ketidakseimbangan kelas, terutama macro-F1 dan balanced accuracy, dibanding baseline, dengan dampak paling jelas pada kelas minoritas seperti Viral Pneumonia yang sering menjadi sumber kesalahan karena kemiripan temuan dengan kategori opacity. Analisis confusion matrix menegaskan bahwa kesalahan utama tetap berasal dari pertukaran antarkelas yang radiografis beririsan, khususnya Viral Pneumonia yang tertukar menjadi Lung Opacity, namun pada model berbasis mask, kesalahan yang tidak relevan secara klinis cenderung berkurang sehingga pola kesalahan lebih terkonsentrasi pada batas pemisahan yang memang sulit. Dari sisi perilaku, audit saliency-in-lung ratio menunjukkan M1 dan M2 lebih konsisten memusatkan atensi pada area paru, serta uji gangguan terbatas area non-paru memperlihatkan M1 dan M2 lebih stabil terhadap perubahan di luar paru sementara M0 mengalami penurunan lebih besar, sehingga kombinasi SILR dan uji gangguan mendukung bukti bahwa pembatasan ROI mengurangi ketergantungan pada petunjuk non-paru yang berpotensi menjadi shortcut. Analisis kalibrasi juga menunjukkan temperature scaling meningkatkan kesesuaian confidence dengan akurasi empiris, dan selective prediction mengindikasikan sistem dapat menahan prediksi pada kasus tidak yakin, yang lebih sesuai untuk layanan kesehatan. Secara keseluruhan, pembatasan pembelajaran pada wilayah paru melalui masking langsung maupun multi-task merupakan strategi data-centric yang efektif untuk meningkatkan kinerja yang lebih merata, memperbaiki auditabilitas, dan meningkatkan robustness terhadap variasi non-klinis, dengan catatan bahwa validasi lokal pada data rumah sakit tetap diperlukan sebelum penggunaan operasional.

REFERENCES

- [1] L. Wang, Z. Q. Lin, dan A. Wong, "COVID Net: a tailored deep convolutional neural network design for detection of COVID 19 cases from chest X ray images," *Scientific Reports*, vol. 10, art. 19549, 2020, doi: 10.1038/s41598-020-76550-z.
- [2] M. E. H. Chowdhury, T. Rahman, A. Khandakar, et al., "Can AI Help in Screening Viral and COVID 19 Pneumonia?," *IEEE Access*, vol. 8, pp. 132665–132676, 2020, doi: 10.1109/ACCESS.2020.3010287.
- [3] M. Roberts, D. Driggs, M. Thorpe, et al., "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID 19 using chest radiographs and CT scans," *Nature Machine Intelligence*, vol. 3, pp. 199–217, 2021, doi: 10.1038/s42256-021-00307-0.
- [4] A. J. DeGrave, J. D. Janizek, dan S. I. Lee, "AI for radiographic COVID 19 detection selects shortcuts over signal," *Nature Machine Intelligence*, vol. 3, pp. 610–619, 2021, doi: 10.1038/s42256-021-00338-7.
- [5] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, dan E. K. Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross sectional study," *PLOS Medicine*, vol. 15, no. 11, e1002683, 2018, doi: 10.1371/journal.pmed.1002683.
- [6] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, dan R. M. Summers, "ChestX ray8: Hospital Scale Chest X ray Database and Benchmarks on Weakly Supervised Classification and Localization of Common Thorax Diseases," dalam *Proc. IEEE CVPR*, 2017, pp. 3462–3471, doi: 10.1109/CVPR.2017.369.
- [7] J. Irvin, P. Rajpurkar, M. Ko, et al., "CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison," dalam *Proc. AAAI Conf. on Artificial Intelligence*, vol. 33, no. 1, pp. 590–597, 2019, doi: 10.1609/aaai.v33i01.3301590.
- [8] A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, et al., "MIMIC CXR, a de identified publicly available database of chest radiographs with free text reports," *Scientific Data*, vol. 6, art. 317, 2019, doi: 10.1038/s41597-019-0322-0.
- [9] A. Bustos, A. Pertusa, J. M. Salinas, dan M. de la Iglesia Vayá, "PadChest: A large chest x ray image dataset

- with multi label annotated reports," *Medical Image Analysis*, vol. 66, art. 101797, 2020, doi: 10.1016/j.media.2020.101797.
- [10] S. Candemir, K. Palaniappan, H. A. Akgul, et al., "Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration," *IEEE Transactions on Medical Imaging*, vol. 33, no. 2, pp. 577–590, 2014, doi: 10.1109/TMI.2013.2290491.
 - [11] O. Ronneberger, P. Fischer, dan T. Brox, "U Net: Convolutional Networks for Biomedical Image Segmentation," dalam *MICCAI 2015, LNCS 9351*, pp. 234–241, 2015, doi: 10.1007/978-3-319-24574-4_28.
 - [12] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, dan D. Batra, "Grad CAM: Visual Explanations from Deep Networks via Gradient Based Localization," dalam *Proc. IEEE ICCV*, 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.
 - [13] C. Guo, G. Pleiss, Y. Sun, dan K. Q. Weinberger, "On Calibration of Modern Neural Networks," dalam *Proc. ICML, PMLR 70*, pp. 1321–1330, 2017.
 - [14] Y. Gal dan Z. Ghahramani, "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning," dalam *Proc. ICML, PMLR 48*, pp. 1050–1059, 2016.
 - [15] T. Rahman, A. Khandakar, Y. Qiblawey, et al., "Exploring the effect of image enhancement techniques on COVID 19 detection using chest X ray images," *Computers in Biology and Medicine*, vol. 132, art. 104319, 2021, doi: 10.1016/j.compbimed.2021.104319.
 - [16] L. Wynants, B. Van Calster, G. S. Collins, et al., "Prediction models for diagnosis and prognosis of covid 19: systematic review and critical appraisal," *BMJ*, vol. 369, m1328, 2020, doi: 10.1136/bmj.m1328.
 - [17] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, dan K. H. Maier-Hein, "nnU Net: a self configuring method for deep learning based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, 2021, doi: 10.1038/s41592-020-01008-z.
 - [18] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, dan J. Liang, "UNet plus plus: A Nested U Net Architecture for Medical Image Segmentation," dalam *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, LNCS 11045*, 2018, pp. 3–11, doi: 10.1007/978-3-030-00889-5_1.
 - [19] O. Oktay, J. Schlemper, L. Le Folgoc, et al., "Attention U Net: Learning Where to Look for the Pancreas," *arXiv:1804.03999*, 2018.
 - [20] X. Liu, S. C. Rivera, D. Moher, et al., "Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT AI extension," *Nature Medicine*, vol. 26, no. 9, pp. 1364–1374, 2020, doi: 10.1038/s41591-020-1034-x.
 - [21] S. C. Rivera, X. Liu, D. Moher, et al., "Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT AI extension," *Nature Medicine*, vol. 26, no. 9, pp. 1351–1363, 2020, doi: 10.1038/s41591-020-1037-7.
 - [22] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, pp. 206–215, 2019, doi: 10.1038/s42256-019-0048-x.
 - [23] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, dan B. Kim, "Sanity Checks for Saliency Maps," dalam *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.
 - [24] R. Geirhos, J. H. Jacobsen, C. Michaelis, et al., "Shortcut learning in deep neural networks," *Nature Machine Intelligence*, vol. 2, no. 11, pp. 665–673, 2020, doi: 10.1038/s42256-020-00257-z.
 - [25] B. Vasey, M. Nagendran, B. Campbell, et al., "Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE AI," *BMJ*, vol. 377, e070904, 2022, doi: 10.1136/bmj-2022-070904.