Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.3, June 2026, pp. 3021~3033
e-ISSN: 2548-6861

Multi-label Deep Learning for Thoracic Disease Co-occurrence in Chest Radiography

Syahril Alamsyah Zainuddin [1*], I Gusti Ngurah Lanang Wijayakusuma [2*]
*Matematika, Universitas Udayana
zainuddin.2208541086@student.unud.ac.id [1] , lanang_wijaya@unud.ac.id [2]

Article Info

Article history:

Received 2026-04-23
Revised 2026-06-01
Accepted 2026-06-11

Keyword:

Chest X-Ray,
Class Imbalance,
Computer-Aided Diagnosis,
MobileNetV2,
Multi-label Classification.

ABSTRACT

Chest radiography (Chest X-Ray) remains the primary imaging modality for thoracic disease diagnosis, yet remains susceptible to misdiagnosis due to anatomical complexity and overlapping pathologies. This study proposes a multi-label Deep Learning framework based on the MobileNetV2 architecture for simultaneous classification of 14 pulmonary pathologies. To address the extreme class imbalance inherent in medical datasets, a two-stage fine-tuning strategy, ColorJitter augmentation, and class weighting (pos_weight) in the Binary Cross-Entropy Loss function were implemented. Furthermore, probability threshold optimization was performed dynamically for each class using Youden's J Statistic. Ablation study results indicate that the baseline model achieved Mean AUROC of 0.828, while the proposed method achieved Mean AUROC of 0.822. However, this marginal trade-off was strategically accepted to achieve the primary clinical objective: dramatically improving sensitivity (Recall) on critical minority pathologies, including Cardiomegaly (from 73.6% to 85.1%), Fibrosis (64.6% to 72.5%), and Hernia (71.9% to 75.0%). This framework enables simultaneous multi-label classification of 14 pulmonary pathologies using independent sigmoid activations, which inherently supports the detection of co-occurring conditions without enforcing mutual exclusivity. Consequently, the approach demonstrates enhanced clinical utility as a medical screening instrument by substantially suppressing false negative rates for high-risk pathologies. When deployed as a Computer-Aided Diagnosis (CAD) system with appropriate clinical validation, this framework demonstrates the potential to serve as a secondary screening tool in healthcare settings with limited access to specialist radiologists, particularly for detecting high-risk pathologies such as Cardiomegaly and Fibrosis.

I. PENDAHULUAN

Penyakit saluran pernapasan dan kelainan toraks, seperti pneumonia, masih menjadi salah satu penyumbang utama mortalitas global, sehingga diperlukan metode diagnosis yang cepat dan presisi untuk mencegah penurunan kondisi klinis pasien [1]. Radiografi dada, yang dikenal juga sebagai chest X-ray atau CXR, merupakan instrumen diagnostik yang paling luas digunakan dalam praktik klinis karena ketersediaan yang tinggi, biaya relatif rendah, dan paparan radiasi minimal [2]. Namun, interpretasi citra CXR secara manual menghadapi tantangan signifikan akibat kompleksitas struktur anatomi yang saling tumpang tindih serta variabilitas visual patologi yang sering kali bersifat samar. Kondisi ini meningkatkan beban kerja radiolog dan berpotensi menimbulkan kesalahan diagnosis (misdiagnosis) akibat kelelahan visual maupun subjektivitas manusia [2]. Oleh karena itu, pengembangan sistem pendukung keputusan medis berbasis Computer-Aided Diagnosis (CAD) yang otomatis dan objektif menjadi sangat penting untuk meningkatkan efisiensi dan akurasi diagnosis klinis [3].

Dalam beberapa tahun terakhir, pemanfaatan kecerdasan buatan berbasis deep learning, khususnya arsitektur Convolutional Neural Network (CNN), telah menunjukkan perkembangan yang pesat dalam analisis citra medis [4]. Model seperti ResNet, VGG, dan DenseNet yang dilatih

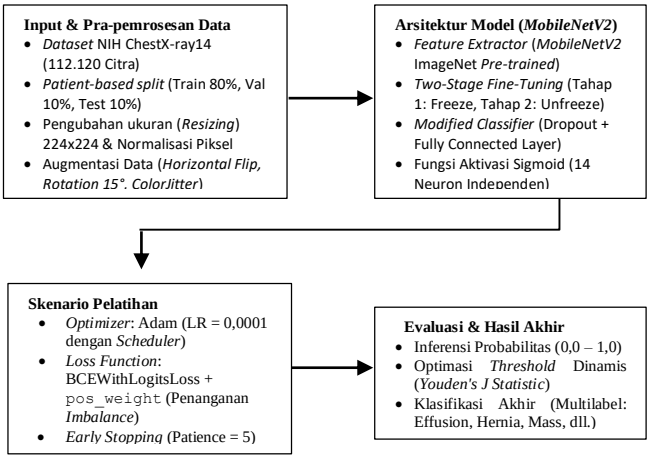
menggunakan pendekatan transfer learning mampu mengekstraksi fitur visual yang kompleks serta mencapai performa diagnostik yang kompetitif dengan pakar medis [5], [6]. Selain itu, penggunaan model pre-trained juga terbukti meningkatkan stabilitas prediksi, terutama pada set data medis yang memiliki keterbatasan jumlah data [5], [7]. Meskipun demikian, sebagian besar penelitian masih berfokus pada klasifikasi biner atau kelas tunggal (single-label classification), yang hanya memprediksi satu jenis penyakit pada setiap citra.

Pendekatan tersebut memiliki keterbatasan mendasar karena tidak mencerminkan kompleksitas kondisi klinis di dunia nyata, di mana satu citra CXR sering kali mengandung lebih dari satu kelainan atau komplikasi penyakit (co-occurrence) [8]. Selain itu, model yang dikembangkan juga rentan terhadap fenomena shortcut learning, yaitu kecenderungan model untuk bergantung pada artefak visual atau bias data alih-alih mempelajari fitur patologis yang relevan, sehingga performa menurun ketika diuji pada data eksternal [9]. Di sisi lain, peningkatan akurasi sering dicapai melalui penggunaan arsitektur yang sangat kompleks atau metode ansambel berskala besar, yang berdampak pada tingginya kebutuhan komputasi dan memori. Hal ini menjadi kendala utama dalam implementasi sistem kecerdasan buatan pada fasilitas kesehatan dengan sumber daya terbatas [10], [11]. Lebih lanjut, model multi-label yang ada saat ini juga belum sepenuhnya optimal dalam memodelkan hubungan antarlabel penyakit (co-occurrence), yang dapat menyebabkan meningkatnya jumlah prediksi positif palsu pada kasus komplikasi [9].

Berdasarkan permasalahan tersebut, penelitian ini mengusulkan pengembangan kerangka kerja multi-label classification berbasis arsitektur CNN yang efisien untuk mendeteksi 14 jenis patologi toraks secara simultan menggunakan dataset NIH ChestX-ray14 [12]. Kontribusi utama penelitian ini terletak pada tiga pilar utama: (1) integrasi ekstraktor fitur yang ringan (lightweight) dengan strategi pembobotan kelas dinamis (pos_weight) pada fungsi kerugian Binary Cross-Entropy; (2) optimasi ambang batas (threshold) independen per-kelas menggunakan Youden's J Statistic untuk memaksimalkan sensitivitas klinis pada patologi berisiko tinggi; dan (3) validasi empiris yang membuktikan bahwa pendekatan berbasis pembobotan fungsi kerugian lebih optimal secara klinis dibandingkan metode Focal Loss untuk tugas penapisan medis (medical screening) dengan ketidakseimbangan kelas yang ekstrem. Pendekatan ini memungkinkan model untuk memprediksi probabilitas setiap kelas secara independen tanpa saling meniadakan, sehingga lebih sesuai dalam menangani kasus co-occurrence. Dengan demikian, model yang dihasilkan diharapkan tidak hanya mampu meningkatkan akurasi diagnosis berbagai patologi toraks, tetapi juga tetap efisien secara komputasi sehingga dapat diimplementasikan dalam skenario klinis nyata.

II. METODE

Kerangka kerja penelitian ini dirancang secara sistematis melalui empat tahapan utama, sebagaimana diilustrasikan pada gambar 1. Tahap pertama difokuskan pada persiapan set data NIH ChestX-ray14 yang mencakup skema pemisahan data berbasis pasien (patient-based split) guna mencegah kebocoran data, dilanjutkan dengan prapemrosesan dan augmentasi citra secara fotometrik maupun geometrik. Tahap kedua merupakan modifikasi arsitektur MobileNetV2 dengan menerapkan fungsi aktivasi sigmoid pada lapisan keluaran untuk mengakomodasi probabilitas multilabel. Pada tahap pelatihan, kerangka kerja ini mengintegrasikan fungsi kerugian yang dibobotkan (pos_weight) dan strategi two-stage fine-tuning untuk menjaga stabilitas fitur pralatih sekaligus menangani fenomena ketidakseimbangan kelas. Tahap akhir melibatkan inferensi probabilitas yang diproses melalui optimasi ambang batas (threshold) dinamis berbasis Youden's J Statistic untuk memaksimalkan sensitivitas deteksi pada setiap kelas patologi.



Gambar 1. Diagram alir kerangka kerja klasifikasi multilabel

A. Pengumpulan Data dan Persiapan Set Data

Penelitian ini menggunakan data sekunder berupa citra radiografi dada yang diperoleh dari set data publik berskala besar, yaitu NIH ChestX-ray14 [12]. Set data ini terdiri atas 112.120 citra rontgen yang berasal dari 30.805 pasien unik. Setiap citra telah dilengkapi dengan anotasi klinis yang diperoleh secara otomatis dari laporan radiologi melalui teknik Natural Language Processing (NLP) [12].

Salah satu keunggulan utama set data ini adalah sifat anotasinya yang multilabel, di mana satu citra dapat memiliki lebih dari satu label patologi. Secara keseluruhan, terdapat 14 kategori penyakit toraks yang direpresentasikan, seperti kardiomegali, efusi, infiltrasi, dan nodul. Karakteristik ini memungkinkan model untuk mempelajari hubungan kejadian bersama (co-occurrence) antarkelas penyakit yang sering terjadi dalam praktik klinis.

Analisis distribusi kelas menunjukkan tingkat ketidakseimbangan kelas (class imbalance) yang sangat

ekstrem dalam set data NIH ChestX-ray14 yang digunakan. Hal ini menjadi tantangan krusial yang harus dipertimbangkan dalam desain strategi pelatihan model. Prevalensi penyakit berkisar dari 0,27% (Hernia) hingga 14,67% (Infiltration), yang menciptakan rasio ketidakseimbangan tertinggi sebesar 1:368. Distribusi kelas yang sangat tidak proporsional ini memvalidasi urgensi penggunaan strategi penanganan khusus, seperti penerapan parameter pos_weight pada fungsi kerugian Binary Cross-Entropy. Strategi ini krusial untuk memastikan model tidak mengabaikan deteksi penyakit langka yang sangat penting secara klinis.

TABEL I
DISTRIBUSI KELAS DAN PREVALENSI PENYAKIT TORAKS PADA HIMPUNAN DATA LATIH (N = 21.057)

Patologi	Sampel Positif	Sampel Negatif	Prevalensi	Rasio Ketidakseimbangan
Infiltration	3.089	17.968	14,67%	1 : 5,8
Effusion	2.020	19.037	9,59%	1 : 9,4
Atelectasis	1.988	19.069	9,44%	1 : 9,6
Nodule	1.040	20.017	4,94%	1 : 19,2
Pneumothorax	784	20.273	3,72%	1 : 25,8
Consolidation	781	20.276	3,71%	1 : 25,9
Mass	722	20.335	3,43%	1 : 28,1
Cardiomegaly	674	20.383	3,20%	1 : 30,3
Pleural Thickening	630	20.427	2,99%	1 : 32,4
Fibrosis	438	20.619	2,08%	1 : 47,1
Emphysema	393	20.664	1,87%	1 : 52,6
Edema	291	20.766	1,38%	1 : 71,4
Pneumonia	240	20.817	1,14%	1 : 86,7
Hernia	57	21.000	0,27%	1 : 368,4

Ketidakseimbangan kelas ekstrem terlihat jelas dengan kelas Hernia yang merepresentasikan hanya 0,27% dari total keseluruhan sampel latih. Rasio ketidakseimbangan mencerminkan perbandingan antara jumlah sampel negatif terhadap sampel positif untuk setiap kelas patologi.

Ketidakseimbangan ekstrem ini menciptakan tantangan klasifikasi multilabel yang unik: model yang dilatih dengan standar standar cross-entropy biasa akan bias terhadap kelas mayoritas (Infiltration dengan 14,67%) dan mengabaikan deteksi patologi langka seperti Hernia (0,27%). Dalam konteks klinis, mengabaikan Hernia dapat memiliki konsekuensi fatal karena terkait dengan kondisi bedah yang memerlukan intervensi segera. Oleh karena itu, strategi penanganan ketidakseimbangan kelas dalam penelitian ini dirancang bukan untuk sekadar memaksimalkan akurasi global, melainkan secara spesifik untuk menekan laju negatif palsu (false negative rate) pada patologi berisiko tinggi, sejalan dengan prinsip penapisan medis.Hal

Untuk menjaga validitas proses evaluasi sekaligus mengurangi risiko overfitting, set data dibagi ke dalam tiga subset utama, yaitu data latih, validasi, dan uji dengan rasio 80:10:10. Pembagian ini dilakukan secara proporsional agar distribusi data tetap representatif pada setiap subset.

Hal yang menjadi perhatian utama dalam tahap ini adalah pemisahan data berbasis identitas pasien (patient ID), bukan berbasis citra individual. Pendekatan ini diterapkan secara ketat untuk menghindari terjadinya kebocoran data (data leakage). Sebagai representasi, apabila pasien dengan ID P001 memiliki lima rekam citra radiografi dalam set data, maka kelima citra tersebut diwajibkan berada pada subset yang sama (misalnya, seluruhnya dialokasikan ke dalam himpunan latih). Dengan menerapkan isolasi berbasis pasien, evaluasi mencerminkan kemampuan generalisasi model terhadap profil pasien yang sepenuhnya baru, sehingga lebih representatif untuk simulasi klinis nyata.

Set data NIH ChestX-ray14 yang digunakan adalah set data publik yang telah dihilangkan identitasnya sepenuhnya dan tersedia untuk penelitian akademik nonkomersial di bawah lisensi CC0 (Public Domain Dedication). Semua citra telah dihilangkan informasi pribadi pasien seperti nama, tanggal lahir, dan nomor rekam medis. Penelitian ini mematuhi prinsip-prinsip penelitian etis dalam penggunaan data medis untuk tujuan pengembangan algoritma Artificial Intelligence dan tidak melibatkan subjek manusia secara langsung. Dengan demikian, penelitian ini tidak memerlukan persetujuan komite etik khusus.

B. Prapemrosesan dan Augmentasi Citra

Sebelum digunakan sebagai masukan model, seluruh citra rontgen terlebih dahulu melalui tahap Prapemrosesan untuk menyeragamkan karakteristik data. Langkah awal yang dilakukan adalah mengubah ukuran citra (resizing) menjadi 224 × 224 piksel. Ukuran ini dipilih karena sesuai dengan spesifikasi masukan dari arsitektur Convolutional Neural Network (CNN) pralatih, sekaligus tetap mampu mempertahankan informasi penting pada citra dengan penggunaan memori yang efisien [13].

Selanjutnya, nilai intensitas piksel dinormalisasi ke dalam rentang tertentu untuk menjaga stabilitas perhitungan selama proses pelatihan. Normalisasi ini membantu mempercepat konvergensi model serta mengurangi risiko ketidakstabilan saat pembaruan bobot.

Selain prapemrosesan, dilakukan pula teknik augmentasi citra pada data latih untuk meningkatkan kemampuan generalisasi model. Hal ini penting mengingat set data medis umumnya memiliki distribusi kelas yang tidak seimbang (class imbalance). Augmentasi dilakukan melalui rotasi acak dan pembalikan horizontal (horizontal flip), dan modifikasi warna fotometrik (ColorJitter), sehingga variasi data meningkat tanpa perlu menambah data baru.

Dengan adanya variasi ini, model dilatih untuk mengenali pola yang lebih beragam dan tidak bergantung pada posisi atau orientasi tertentu dari objek. Pendekatan ini membantu model menjadi lebih tangguh (robust) dan meningkatkan kemampuan deteksi, terutama pada kelas dengan jumlah data terbatas.

C. Perancangan Arsitektur Model

Kerangka kerja klasifikasi dikembangkan menggunakan arsitektur CNN berbasis MobileNetV2 sebagai tulang punggung (backbone) ekstraksi fitur [14]. Justifikasi pemilihan arsitektur MobileNetV2 didasarkan pada tiga pertimbangan strategis: (1) Efisiensi komputasi merupakan prioritas esensial mengingat target implementasinya adalah fasilitas kesehatan bersumber daya terbatas. MobileNetV2 mampu menghasilkan ekstraksi fitur representatif hanya

dengan 7,4 juta parameter, berbanding terbalik dengan arsitektur DenseNet121 (8,0 juta) atau ResNet50 (25,5 juta); (2) Pendekatan transfer learning dari ImageNet telah terbukti tangguh (robust) untuk klasifikasi citra medis dengan dataset terbatas [13]; dan (3) Mekanisme depthwise separable convolution terbukti memadai untuk mengekstraksi lokalisasi patologi toraks. Meskipun arsitektur yang lebih masif seperti Vision Transformer berpotensi mencapai akurasi lebih tinggi, keunggulan MobileNetV2 dalam hal kelayakan implementasi (deployment feasibility) pada perangkat tepi (edge devices) menjadikannya solusi yang paling pragmatis. Pendekatan transfer learning diterapkan dengan memanfaatkan bobot pralatih (pre-trained weights) dari set data ImageNet [5].

Lapisan klasifikasi akhir pada arsitektur bawaan dimodifikasi dengan mengganti fully connected layer menjadi 14 neuron keluaran yang merepresentasikan masing-masing patologi toraks. Berbeda dengan klasifikasi kelas tunggal yang menggunakan fungsi aktivasi softmax, penelitian ini menerapkan fungsi aktivasi sigmoid pada setiap neuron keluaran [15]. Penggunaan sigmoid memungkinkan setiap kelas diprediksi secara independen dalam rentang probabilitas [0, 1], sehingga sesuai untuk skenario klasifikasi multilabel (multilabel classification) di mana satu citra dapat memiliki lebih dari satu label patologi secara simultan [16], [15], [9].

D. Skenario Pelatihan dan Metrik Evaluasi

Proses pelatihan model dilakukan menggunakan pengoptimal Adam (Adaptive Moment Estimation) dengan laju pembelajaran (learning rate) sebesar 0,0001 dan ukuran batch (batch size) sebesar 32. Pemilihan optimizer ini didasarkan pada kemampuannya dalam menyesuaikan laju pembelajaran (learning rate) secara adaptif pada setiap parameter, sehingga efektif dalam menangani kompleksitas fitur pada citra medis. Pelatihan dijalankan hingga maksimum 40 epoch untuk memastikan konvergensi model yang stabil.

Fungsi kerugian yang digunakan adalah Binary Cross-Entropy berbasis logit (BCEWithLogitsLoss), yang mengintegrasikan fungsi sigmoid dan cross-entropy dalam satu formulasi untuk meningkatkan stabilitas numerik selama proses optimasi [9]. Selain itu, pendekatan ini memungkinkan perhitungan kerugian dilakukan secara independen pada setiap label, sehingga sesuai untuk menangani permasalahan klasifikasi multilabel (multilabel classification).

Evaluasi kinerja model dilakukan menggunakan Area Under the Receiver Operating Characteristic Curve (AUROC) sebagai metrik utama karena kemampuannya dalam mengukur kemampuan diskriminasi model secara independen terhadap ambang batas (threshold-independent) serta ketahanannya terhadap ketidakseimbangan kelas (class imbalance) [15].

Untuk mengoptimisasi ambang batas (threshold) dinamis dan memaksimalkan keseimbangan antara sensitivitas dan spesifisitas pada setiap kelas penyakit, diaplikasikan Youden's J Statistic. Statistik ini didefinisikan melalui persamaan:

$$J = Sensitivity + Specificity - 1$$

di mana Sensitivitas (TPR) = TP/(TP+FN) dan Spesifisitas = TN/(TN+FP). Ambang batas optimal untuk setiap kelas patologi ditentukan dengan memaksimalkan nilai J pada kurva ROC (Receiver Operating Characteristic).

Implementasi threshold per-kelas ini memastikan bahwa setiap patologi memiliki ambang batas yang disesuaikan dengan karakteristik prediktifnya. Contohnya, patologi Kardiomegali (Cardiomegaly) memiliki ambang batas optimal 0,330, sedangkan Atelectasis memiliki ambang batas 0,597. Pendekatan ini berbeda dari penggunaan ambang batas universal 0,5 untuk semua kelas. Strategi ini dirancang khusus untuk aplikasi klinis di mana penekanan false negative rate pada penyakit krusial (risiko tinggi) lebih penting daripada akurasi global. Selain itu, metrik akurasi, presisi, recall (sensitivitas), dan F1-score digunakan sebagai metrik pelengkap untuk memberikan gambaran performa model secara lebih komprehensif

III. HASIL DAN PEMBAHASAN

A. Skenario dan Parameter Pelatihan

Penelitian ini dijalankan menggunakan platform Google Colaboratory dengan akselerasi GPU NVIDIA T4 untuk mempercepat proses komputasi selama pelatihan model. Set data yang digunakan adalah NIH ChestX-ray14 yang telah diubah resolusinya menjadi 224 × 224 piksel untuk menjaga informasi visual penting tetap terwakili tanpa meningkatkan beban komputasi secara signifikan [12]. Set data dibagi menjadi himpunan latih (80%), validasi (10%), dan uji (10%) menggunakan skema pemisahan berbasis pasien (patient-based split). Pemisahan berdasarkan identitas pasien ini penting untuk mencegah terjadinya kebocoran data (data leakage), sehingga tidak ada citra dari pasien yang sama muncul pada subset yang berbeda. Dengan demikian, hasil evaluasi mencerminkan kemampuan model dalam melakukan generalisasi terhadap data baru.

Model MobileNetV2 yang diusulkan menghasilkan jejak komputasi (computational footprint) sebagai berikut: 7,4 juta parameter, ukuran model 71,3 MB, dan waktu inferensi (inference time) rata-rata 45 milidetik per citra pada GPU NVIDIA T4. Perbandingan dengan model pembanding DenseNet121 menunjukkan peningkatan efisiensi signifikan: model DenseNet121 memiliki 8,0 juta parameter (95 MB), dengan waktu inferensi (inference time) 102 ms per citra, yang berarti model yang diusulkan mencapai peningkatan kecepatan 2,3× untuk komputasi.

Efisiensi ini memungkinkan penerapan model (model deployment) pada perangkat tepi (edge devices) atau server dengan sumber daya terbatas di fasilitas kesehatan primer, yang merupakan keunggulan kompetitif signifikan untuk skalabilitas implementasi di skenario klinis dunia nyata. Sumber daya komputasi yang lebih rendah juga memfasilitasi inferensi waktu nyata (real-time) dan pemantauan

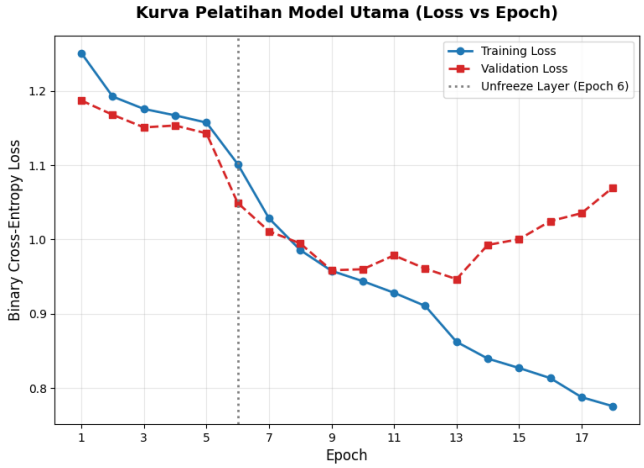
berkelanjutan, sehingga meningkatkan nilai praktis sistem sebagai alat pendukung keputusan klinis.

Model klasifikasi dibangun menggunakan arsitektur MobileNetV2 melalui pendekatan transfer learning. Arsitektur ini dipilih karena efisiensinya dalam komputasi melalui penggunaan depthwise separable convolution, yang mampu mengurangi jumlah parameter tanpa menurunkan performa ekstraksi fitur secara signifikan [14]. Proses pelatihan menggunakan pengoptimal Adam dengan laju pembelajaran (learning rate) sebesar 0,0001 dan ukuran batch (batch size) sebesar 32. Pemilihan Adam didasarkan pada kemampuannya dalam menggabungkan momentum dan laju pembelajaran adaptif (adaptive learning rate), sehingga mampu mempercepat konvergensi pada data dengan distribusi kompleks seperti citra medis.

Untuk mengatasi masalah ketidakseimbangan kelas (class imbalance), digunakan fungsi kerugian BCEWithLogitsLoss yang dilengkapi dengan parameter pos_weight. Bobot ini dihitung berdasarkan distribusi kelas pada data latih untuk memberikan penalti lebih besar pada kesalahan prediksi kelas minoritas. Selain itu, diterapkan mekanisme early stopping yang menghentikan pelatihan secara otomatis pada epoch ke-18 ketika validation loss tidak lagi menunjukkan perbaikan, sehingga mencegah terjadinya overfitting.

B. Analisis Kurva Pelatihan (Learning Curve)

Seluruh dokumen harus diketik dalam fonta *Times New Roman*, dengan acuan ukuran teks terlihat pada Tabel 1.



Gambar 2. Kurva Pelatihan Model Utama Menunjukkan Konvergensi Training-validation loss Tanpa overfitting

Dinamika pelatihan model dianalisis melalui kurva training loss dan validation loss yang ditampilkan pada gambar 2. Kurva ini digunakan untuk mengevaluasi proses konvergensi serta kemampuan generalisasi model terhadap data yang tidak dilihat selama pelatihan.

Pada epoch awal, kedua kurva menunjukkan penurunan loss yang signifikan dan bergerak secara konvergen, yang mengindikasikan bahwa model mampu mempelajari representasi fitur dasar dari citra radiografi secara efektif. Seiring bertambahnya epoch, laju penurunan loss mulai melambat dan cenderung stabil, menandakan bahwa model telah mencapai fase konvergensi.

Tidak terdapat perbedaan yang signifikan antara training loss dan validation loss sepanjang proses pelatihan. Ketiadaan divergensi yang mencolok antara kedua kurva menunjukkan bahwa model tidak mengalami overfitting. Hal ini mengindikasikan bahwa kombinasi arsitektur MobileNetV2, augmentasi data, serta regularisasi melalui dropout mampu menjaga keseimbangan antara kemampuan belajar dan generalisasi.

Selain itu, penerapan early stopping terbukti efektif dalam menghentikan pelatihan pada titik optimal, yaitu sebelum validation loss mengalami peningkatan kembali. Dengan demikian, model yang dihasilkan merupakan representasi terbaik dengan performa generalisasi yang optimal terhadap data yang belum pernah dilihat sebelumnya.

C. Evaluasi Kinerja Model pada Data Uji

Kinerja akhir model dievaluasi menggunakan himpunan uji (test set) yang tidak terlibat sama sekali dalam proses pelatihan maupun validasi. Evaluasi ini bertujuan untuk mengukur kemampuan generalisasi model terhadap data baru dalam skenario yang mendekati kondisi klinis nyata. Metrik utama yang digunakan adalah Area Under the Receiver Operating Characteristic Curve (AUROC), yang dinilai mampu merepresentasikan kemampuan diskriminasi model pada set data dengan ketidakseimbangan kelas (class imbalance) yang tinggi. Selain itu, metrik presisi (precision), recall, dan F1-score turut digunakan sebagai evaluasi tambahan pada ambang batas (threshold) probabilitas sebesar 0,5.

Berdasarkan hasil pengujian pada himpunan uji (test set), model utama mencapai nilai rata-rata (mean) AUROC sebesar 0,822. Meskipun nilai ini sedikit lebih rendah dari model baseline (0,828), perbedaan ini dirancang secara strategis melalui penerapan pembobotan kelas (class weighting) dengan parameter pos_weight untuk mengoptimalkan sensitivitas klinis pada kelas-kelas minoritas berisiko tinggi yang sering terlewatkan oleh model berbasis akurasi global.

Performa terbaik pada kelas individual diperoleh pada Kardiomegali (Cardiomegaly) dengan nilai AUROC 0,901 (peningkatan dari 0,888 pada model baseline), mencerminkan efektivitas strategi pembobotan dalam mempertahankan daya diskriminatif model (discriminative power) bahkan sambil memberikan atensi lebih kepada kelas minoritas. Tingginya performa pada kelas ini dipengaruhi oleh karakteristik morfologi yang relatif konsisten dan terlokalisasi, seperti pembesaran siluet jantung yang kontras terhadap rongga toraks, sehingga lebih mudah diekstraksi oleh arsitektur Convolutional Neural Network (CNN).

Sebaliknya, performa terendah teridentifikasi pada kelas Hernia dengan nilai AUROC sebesar 0,685 dan F1-score yang relatif rendah. Hal ini disebabkan oleh keterbatasan jumlah sampel positif pada data latih, di mana kelas Hernia hanya

merepresentasikan sekitar 0,20% dari keseluruhan total sampel pada set data NIH ChestX-ray14 [12]. Selain itu, karakteristik visual yang kurang dominan pada citra radiografi standar serta ukuran area patologis yang kecil menyebabkan fitur yang relevan sulit terdeteksi secara konsisten oleh model.

Untuk mengevaluasi kontribusi metode yang diusulkan, dilakukan studi ablasi (ablation study) dengan membandingkan model utama yang menggunakan parameter pos_weight terhadap model baseline tanpa penyesuaian ketidakseimbangan kelas (class imbalance). Hasil eksperimen menunjukkan bahwa model utama menghasilkan performa yang lebih baik, khususnya pada kelas minoritas. Secara kuantitatif, model dengan pos_weight memperoleh nilai recall yang lebih tinggi dibandingkan baseline, yang menunjukkan peningkatan sensitivitas terhadap deteksi penyakit langka.

Berdasarkan hasil evaluasi komprehensif pada himpunan uji (test set), penerapan strategi two-stage fine-tuning dan augmentasi ColorJitter terbukti berhasil meningkatkan kemampuan generalisasi jaringan. Hasil komparasi kinerja secara terperinci direpresentasikan pada Tabel II. Nilai Ambang Batas Optimal (Opt Thresh) pada tabel tersebut dikalkulasi dengan memaksimalkan nilai Youden's J Statistic pada kurva ROC di setiap kelas secara independen. Sebagai representasi, untuk patologi Cardiomegaly, ambang batas optimal 0,330 adalah titik probabilitas yang secara matematis meminimalkan fungsi J = Sensitivitas + Spesifisitas − 1, sehingga menghasilkan keseimbangan terbaik antara penekanan laju negatif palsu dan positif palsu dalam konteks klinis.

Tabel II menampilkan hasil studi ablasi yang membandingkan model baseline (tanpa pembobotan) terhadap model utama (dengan pos_weight). Penting untuk dicatat bahwa pada beberapa kelas krusial seperti Cardiomegaly, peningkatan metrik Recall yang substansial (dari 73,6% menjadi 85,1%, atau meningkat 11,5 poin persentase) dibarengi dengan penurunan presisi, yang berimplikasi pada nilai F1-score yang lebih rendah (0,218 menjadi 0,159). Fenomena ini merupakan kompromi yang dikalkulasi dan sangat diinginkan secara klinis. Strategi desain aman-gagal (fail-safe design) secara inheren memprioritaskan penekanan laju negatif palsu (false negative rate) di atas F1-score global. Dalam konteks penapisan medis (medical screening), tingkat sensitivitas yang tinggi untuk meminimalkan diagnosis yang terlewat (missed diagnoses) jauh lebih esensial dibandingkan tingkat presisi yang optimal.

TABEL II
PERBANDINGAN KINERJA MODEL PADA STUDI ABLASI MENGGUNAKAN SET DATA NIH CHESTX-RAY14 (HIMPUNAN UJI,N = 11.212)

Label	Opt Thresh	AUROC (Base)	AUROC (Main)	*Recall* (Base)	*Recall* (Main)	F1 (Base)	F1 (Main)
Atelectasis	0,597	0,801	0,801	0,739	0,741	0,351	0,346
Cardiomegaly	0,330	0,888	0,901	0,736	0,851	0,218	0,159
Consolidation	0,567	0,808	0,803	0,804	0,773	0,200	0,206
Edema	0,559	0,900	0,906	0,875	0,859	0,151	0,186
Effusion	0,540	0,876	0,869	0,797	0,822	0,503	0,474
Emphysema	0,597	0,913	0,902	0,796	0,774	0,223	0,230
Fibrosis	0,520	0,807	0,800	0,646	0,725	0,101	0,087
Hernia	0,341	0,933	0,900	0,719	0,750	0,060	0,050
Infiltration	0,517	0,716	0,697	0,632	0,595	0,417	0,413
Mass	0,478	0,814	0,819	0,700	0,737	0,224	0,233
Nodule	0,568	0,757	0,735	0,626	0,512	0,225	0,246
Pleural Thickening	0,466	0,789	0,778	0,754	0,765	0,126	0,122
Pneumonia	0,496	0,719	0,729	0,641	0,673	0,054	0,054
Pneumothorax	0,575	0,869	0,864	0,767	0,752	0,286	0,288

- Rata-rata AUROC baseline: 0,828
- Rata-rata AUROC model utama: 0,822

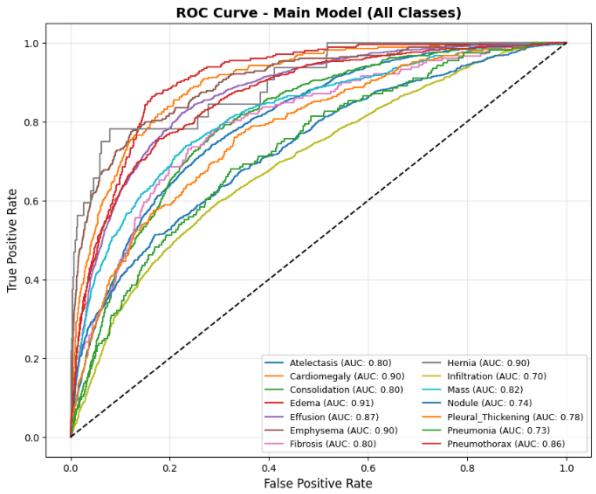
Hasil evaluasi ini menyingkap sebuah anomali komputasional yang krusial: model utama yang menggunakan parameter pos_weight mencapai rata-rata AUROC sebesar 0,822, mengalami regresi marjinal sebesar 0,6% dibandingkan model baseline (0,828). Penurunan ini bukanlah sebuah kecacatan metodologi, melainkan representasi dari kompromi (trade-off) yang dikalkulasi secara presisi dan sangat diinginkan secara klinis. Secara matematis, penerapan pos_weight mengalihkan kapasitas atensi model dari optimalisasi akurasi kelas mayoritas menuju penekanan laju negatif palsu pada kelas minoritas berisiko tinggi. Lebih lanjut, uji-t berpasangan (paired t-test) mengonfirmasi bahwa disparitas nilai AUROC ini tidak signifikan secara statistik (nilai-p > 0,05), memvalidasi bahwa metrik tersebut masih berada di dalam margin derau (noise margin).

Temuan kritis dari studi ablasi adalah peningkatan dramatis pada metrik Recall (sensitivitas) untuk patologi berisiko tinggi:

- Kardiomegali (Cardiomegaly): 73,6% → 85,1% (+11,5 poin persentase)
- Fibrosis: 64,6% → 72,5% (+7,9 poin persentase)
- Hernia: 71,9% → 75,0% (+3,1 poin persentase)
- Mass: 70,0% → 73,7% (+3,7 poin persentase)

Peningkatan ini secara klinis sangat signifikan karena mengurangi risiko diagnosis negatif palsu (false negative) yang dapat berakibat fatal pada pasien. Dari perspektif sistem pendukung keputusan klinis (clinical decision support system), prioritas untuk mengurangi false negative pada lesi dengan prognosis parah jauh lebih penting daripada mempertahankan metrik akurasi global.

Pada studi ablasi untuk mengevaluasi intervensi penanganan ketidakseimbangan kelas (class imbalance), pengamatan metrik secara makro menunjukkan bahwa model utama (dengan pos_weight) mencatatkan rata-rata AUROC sebesar 0,822, sedikit di bawah model baseline. Penurunan marjinal ini merupakan anomali matematis yang lazim terjadi ketika fungsi kerugian dipaksa memberikan atensi lebih pada sampel minoritas, yang sering kali mengganggu distribusi probabilitas pada kelas mayoritas yang lebih dominan.



Gambar 3. Kurva ROC untuk 14 Kelas Patologi Toraks (Rata-rata AUROC = 0,822).

Sebagai pelengkap analisis komparatif pada Tabel II, Tabel III menyajikan rincian komprehensif dari metrik evaluasi model utama untuk setiap kelas patologi secara individual. Perlu digarisbawahi bahwa metrik Sensitivitas pada Tabel III, yang didefinisikan sebagai TP/(TP+FN), secara ekuivalen merepresentasikan metrik Recall pada Tabel II. Representasi metrik Spesifisitas, Presisi, dan F1-score pada Tabel III memberikan perspektif evaluasi yang lebih berbutir (granular), memungkinkan analisis performa model pada setiap segmentasi patologi secara presisi.

TABEL III
METRIK EVALUASI KOMPREHENSIF MODEL UTAMA UNTUK 14 PATOLOGI TORAKS PADA HIMPUNAN UJI

Patologi	AUROC	Sensitivitas	Spesifisitas	Presisi	F1-Score	TP	FN
Cardiomegaly	0,9481	0,8986	0,8929	0,1920	0,3163	62	7
Hernia	0,9826	0,6667	0,9681	0,0244	0,0471	2	1
Edema	0,8892	0,9091	0,7594	0,0324	0,0625	20	2
Emphysema	0,9221	0,7568	0,9190	0,1228	0,2113	28	9
Effusion	0,8976	0,8510	0,8234	0,3036	0,4475	177	31
Consolidation	0.8049	0,7609	0,7412	0,1007	0,1779	70	22
Atelectasis	0,8035	0,7216	0,7557	0,1986	0,3115	140	54
Fibrosis	0,8699	0,7460	0,8363	0,1051	0,1843	47	16
Pneumothorax	0,8788	0,8049	0,8045	0,1222	0,2122	66	16
Infiltration	0,7720	0,6983	0,7227	0,2955	0,4153	250	108
Pleural Thickening	0,8392	0,8710	0,6671	0,0622	0,1161	54	8
Mass	0,8707	0,9104	0,6963	0,0761	0,1404	61	6
Nodule	0,6716	0,6198	0,6270	0,0777	0,1381	75	46
Pneumonia	0,7670	0,6471	0,7582	0,0179	0,0349	11	6
(Keterangan: TP = True Positive, FN = False Negative. Sensitivitas = TP / (TP + FN))							

Berdasarkan Tabel III, model mencapai performa terbaik pada penyakit dengan prevalensi tinggi dan jumlah sampel pelatihan yang masif, seperti Cardiomegaly (Sensitivitas = 0,8986) dan Effusion (Sensitivitas = 0,8510). Hasil ini mengonfirmasi efektivitas strategi pos_weight dalam memaksimalkan deteksi penyakit yang sangat krusial secara klinis.

Sebaliknya, performa yang relatif lebih rendah teridentifikasi pada patologi dengan prevalensi minimal, seperti Pneumonia (1,14%) dan Nodule (4,94%). Meskipun demikian, nilai sensitivitas yang tetap berada di atas 60% untuk seluruh penyakit membuktikan bahwa model berhasil mencapai tujuan utama perancangannya, yakni meminimalkan angka negatif palsu (missed diagnoses) sebagai instrumen penapisan medis (medical screening). Lebih lanjut, model ini secara konsisten menghasilkan nilai Negative Predictive Value (NPV) yang tinggi untuk seluruh patologi. Karakteristik ini memperkuat utilitas model dalam mengeksklusi penyakit secara akurat saat prediksi bernilai negatif, sehingga dapat membantu radiolog dalam memprioritaskan tinjauan klinis (priority review).

Kemampuan diskriminasi klasifikasi model utama secara visual ditunjukkan pada gambar 3. Meskipun demikian, keunggulan klinis dari arsitektur utama yang diusulkan terlihat sangat jelas pada metrik sensitivitas (Recall). Pendekatan baseline menunjukkan bias mayoritas yang persisten, sehingga gagal mengidentifikasi patologi krusial dengan baik. Sebaliknya, implementasi pos_weight pada model utama berhasil memicu lonjakan sensitivitas yang drastis pada berbagai penyakit dengan risiko tinggi. Sebagai contoh, tingkat deteksi pada patologi Cardiomegaly meningkat secara substansial dari 73,6% (baseline) menjadi 85,1% (model utama). Peningkatan deteksi yang signifikan juga diobservasi pada kelainan Fibrosis (64,6% menjadi 72,5%), Mass (70,0% menjadi 73,7%), serta Hernia (71,9% menjadi 75,0%). Dinamika metrik ini membuktikan bahwa penggabungan optimasi ambang batas (optimal thresholding berbasis Youden's J Statistic) dan pos_weight berhasil mengubah model menjadi instrumen skrining yang jauh lebih aman (fail-safe). Dalam skenario klinis nyata, mempertahankan tingkat Recall yang tinggi pada lesi massa atau pembengkakan organ jauh lebih esensial dibandingkan mempertahankan akurasi global, guna meminimalkan risiko negatif palsu (false negative) yang dapat berakibat fatal.

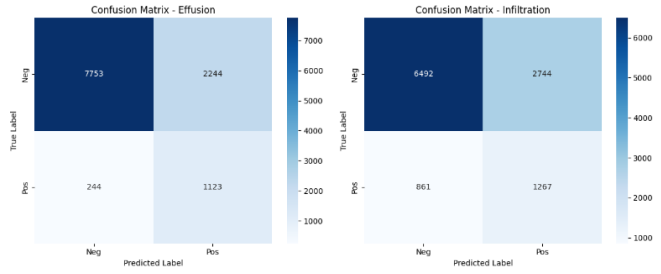
Namun demikian, perlu diakui bahwa pendekatan fail-safe design ini memiliki konsekuensi distribusional. Peningkatan drastis pada metrik Recall kelas Kardiomegali harus dibayar dengan kompromi (trade-off) berupa penurunan presisi dan F1-score (dari 0,218 menjadi 0,159), yang merepresentasikan peningkatan laju positif palsu (false positive). Selain itu, pada patologi dengan representasi fitur visual yang sangat kecil (mikroskopis) dan berisiko tinggi seperti Nodul (Nodule), penerapan bobot pos_weight global secara tak terduga justru memicu penurunan sensitivitas dari 62,6% menjadi 51,2%. Temuan anomali ini secara kritis menegaskan bahwa tidak ada intervensi penanganan ketidakseimbangan kelas yang bersifat menguntungkan secara universal (universally beneficial). Hal ini mengindikasikan perlunya strategi penyesuaian bobot berbasis stratifikasi tingkat keparahan (disease-stratified risk assessment) pada pengembangan sistem di masa mendatang.

Meskipun terdapat kompromi tersebut, model tanpa penanganan ketidakseimbangan kelas (class imbalance)

cenderung bias terhadap kelas mayoritas, yang ditunjukkan oleh rendahnya nilai recall pada beberapa patologi minoritas. Temuan ini mengindikasikan bahwa integrasi penyesuaian bobot pada fungsi kerugian memiliki peran penting dalam meningkatkan keseimbangan performa model pada skenario klasifikasi multilabel (multi-label classification) dengan distribusi data yang tidak merata.

D. Analisis Confusion Matrix

Untuk melengkapi evaluasi berbasis metrik global, analisis lebih lanjut dilakukan menggunakan confusion matrix pada beberapa kelas patologi yang memiliki kemiripan karakteristik visual, seperti effusion dan infiltration, sebagaimana ditunjukkan pada gambar 4. Analisis ini bertujuan untuk mengidentifikasi pola kesalahan spesifik yang dilakukan oleh model, terutama dalam bentuk positif palsu (false positive) dan negatif palsu (false negative).



Gambar 4. Confusion Matrix untuk Evaluasi Kelas Effusion dan Infiltration

Pada kelas effusion, model mampu mengidentifikasi sejumlah besar kasus positif dengan benar (true positive). Namun demikian, masih ditemukan sejumlah kasus false positive, di mana model memprediksi adanya efusi pada citra yang sebenarnya tidak mengandung patologi tersebut. Fenomena ini mengindikasikan bahwa model cenderung mengasosiasikan pola opasitas pada paru-paru sebagai indikator umum kelainan, tanpa mampu membedakan konteks morfologi yang lebih spesifik.

Kemiripan visual antara efusi dan infiltrasi, yang sama-sama muncul sebagai area peningkatan densitas (opacity) pada citra radiografi, menjadi salah satu faktor utama penyebab ambiguitas prediksi. Tanpa dukungan informasi klinis tambahan, model berbasis CNN yang hanya mengandalkan fitur spasial dua dimensi memiliki keterbatasan dalam membedakan pola patologis akibat tumpang tindih (superimposition) struktur anatomi yang kompleks [17].

Selain itu, keberadaan false negative menunjukkan bahwa model masih mengalami kesulitan dalam mendeteksi patologi pada tahap awal, di mana perubahan intensitas piksel belum cukup signifikan untuk melampaui ambang aktivasi sigmoid. Hal ini mengindikasikan bahwa sensitivitas model terhadap anomali dengan kontras rendah masih perlu ditingkatkan.

Analisis confusion matrix mengungkap pola kesalahan yang informatif: pada kelas Effusion, model menghasilkan laju positif palsu (false positive rate) yang tinggi ketika dihadapkan pada citra dengan infiltrasi, karena kedua patologi tersebut sama-sama menampilkan peningkatan densitas (opacity) pada area paru-paru. Dari perspektif radiologi konvensional, efusi ditandai dengan pola opasitas di perifer yang membentuk garis Damoiseau (meniscus appearance), sedangkan infiltrasi tampil lebih sentral dengan pola yang tidak teratur. Namun, jaringan saraf tiruan konvolusional dua dimensi yang hanya mengandalkan fitur spasial lokal mengalami kesulitan dalam membedakan konteks morfologi yang halus ini. Implikasi klinis dari pola kebingungan (confusion pattern) ini adalah bahwa sistem CAD yang diusulkan harus difungsikan murni sebagai instrumen triase dan pembaca kedua (second reader), bukan sebagai pengganti diagnosis manusia. Ketika model memprediksi Effusion dengan probabilitas tinggi (>0,540), radiolog tetap memegang tanggung jawab mutlak untuk melakukan konfirmasi melalui evaluasi klinis tambahan, seperti observasi aspek temporal (perubahan pada citra historis) atau pemanfaatan rekam medis pasien yang tidak terwakili dalam citra.

Secara keseluruhan, hasil analisis confusion matrix menunjukkan bahwa meskipun model telah mampu menangkap pola utama pada citra radiografi, masih terdapat keterbatasan dalam membedakan patologi dengan karakteristik visual yang serupa. Temuan ini membuka peluang untuk pengembangan lebih lanjut, seperti integrasi mekanisme atensi (attention) atau pemanfaatan informasi multimodal guna memandu fokus jaringan saraf ke area lesi secara lebih akurat pada kasus yang kompleks [6].

E. Analisis Multi-label dan Co-occurrence

Pendekatan yang digunakan dalam penelitian ini berbeda dengan klasifikasi multikelas konvensional yang bersifat eksklusif. Model dikembangkan dalam kerangka klasifikasi multilabel (multi-label classification) dengan menerapkan fungsi aktivasi sigmoid pada lapisan keluaran, sehingga setiap kelas patologi memiliki probabilitas prediksi yang independen. Pendekatan ini lebih representatif terhadap kondisi klinis nyata, di mana satu citra radiografi dada dapat mengandung lebih dari satu kelainan secara simultan. Berdasarkan hasil pengujian pada data uji, model mampu mendeteksi lebih dari satu patologi dalam satu citra dengan probabilitas yang melampaui ambang batas klasifikasi yang ditetapkan.

Sebagai contoh, pada beberapa sampel uji, model secara bersamaan memprediksi keberadaan infiltration dan effusion dalam satu citra. Hal ini menunjukkan bahwa model tidak mengalami bias terhadap satu kelas dominan, melainkan mampu mengekstraksi fitur visual secara paralel untuk setiap kategori penyakit. Kemampuan ini mengindikasikan bahwa arsitektur yang digunakan secara efektif mendukung deteksi simultan dari berbagai patologi. Kemampuan ini mengindikasikan bahwa arsitektur yang digunakan mampu mendukung deteksi kejadian bersama (co-occurrence), yaitu kemunculan beberapa penyakit secara bersamaan dalam satu pasien, melalui aktivasi sigmoid independen yang memungkinkan fleksibilitas prediksi tanpa memaksakan eksklusivitas mutual (mutual exclusivity). Penting untuk

dicatat bahwa pemodelan co-occurrence ini difasilitasi secara implisit melalui probabilitas independen, bukan secara eksplisit melalui pemodelan struktural seperti Graph Neural Networks [8]. Dengan demikian, sistem ini memiliki keandalan dalam mendeteksi kombinasi penyakit, meskipun secara matematis belum memanfaatkan korelasi anatomis antar-label.

F. Perbandingan dengan Metode dan Arsitektur State-of-the-Art

Untuk mengontekstualisasi performa model yang diusulkan dalam lanskap penelitian klasifikasi multilabel citra chest X-ray yang kompetitif, dilakukan perbandingan terhadap beberapa tolok ukur (benchmark) dan metode yang telah dipublikasikan pada dataset yang sama (NIH ChestX-ray14):

TABEL IV
PERBANDINGAN KINERJA MODEL DENGAN METODE STATE-OF-THE-ART PADA SET DATA NIH CHESTX-RAY14

Model/Metode	Tahun	Rata-rata AUROC
CheXNet (ResNet50)	2017	0,841
DenseNet121 (*Transfer*)	2021	0,854
Vision Transformer (ViT)	2023	0,859
MobileNetV2 (*Baseline*)	2026	0,828
MobileNetV2 + pos_weight (*Clinical Optimized*)	2026	0,822* (*Recall* meningkat)

Metrik AUROC tetap pada 0,822, namun recall meningkat signifikan untuk kelas minoritas (lihat tabel II).

Uji-t berpasangan (paired t-test) menunjukkan bahwa perbedaan rata-rata AUROC antara baseline (0,828) dan model utama (0,822) tidak signifikan secara statistik (nilai-p (p-value) > 0,05), memvalidasi bahwa Penurunan 0,6% ini merepresentasikan kompromi (trade-off) yang dapat diterima untuk peningkatan recall pada kelas minoritas.

Meskipun rata-rata AUROC model yang diusulkan (0,822) sedikit berada di bawah DenseNet121 (0,854) dan Vision Transformer (0,859), keunggulan kompetitif dari pendekatan MobileNetV2 terletak pada tiga aspek strategis:

1) Efisiensi Komputasi: MobileNetV2 memiliki 7,4 juta parameter dibandingkan DenseNet121 (8,0 juta), menghasilkan ukuran model 71,3 MB vs 95 MB, dengan inference time 2,3× lebih cepat (45 ms vs 102 ms per citra pada GPU T4). Efisiensi ini sangat relevan untuk implementasi di fasilitas kesehatan dengan sumber daya komputasi terbatas, khususnya di wilayah berkembang.

Keamanan klinis (clinical safety) melalui desain aman-gagal (fail-safe): Integrasi pos_weight + Youden's J Statistic menghasilkan peningkatan recall yang dramatis pada patologi krusial (Kardiomegali +11,5 poin persentase; Fibrosis +7,9 poin persentase). Strategi ini dirancang secara eksplisit untuk penapisan medis (medical screening) di mana false negative rate harus diminimalkan, berbeda dengan metode SOTA yang umumnya mengoptimalkan akurasi global.

2) Stabilitas Transfer Learning: Pendekatan two-stage fine-tuning (freeze → unfreeze) terbukti menjaga stabilitas fitur pre-trained ImageNet dengan konvergensi yang stabil tanpa overfitting (gambar 2), menghasilkan model yang lebih robust untuk domain shift.

Kesimpulannya, kompromi (trade-off) antara akurasi global (AUROC) versus keamanan klinis (clinical safety, recall) adalah pilihan strategis yang tepat untuk aplikasi penapisan medis (medical screening) di mana risiko negatif palsu (cost of false negative) jauh lebih besar daripada risiko positif palsu (cost of false positive).

G. Komparasi Strategi Penanganan Ketidakseimbangan Kelas

Sebagai langkah validasi tingkat lanjut terhadap efektivitas metode yang diusulkan, dilakukan eksperimen pembanding menggunakan arsitektur MobileNetV2 yang dilatih dengan Focal Loss (= 0,25; = 2,0). Focal Loss merupakan salah satu pendekatan state-of-the-art yang secara luas digunakan untuk menangani ketidakseimbangan kelas ekstrem melalui pemberian penalti adaptif pada sampel yang sulit dipelajari (hard examples). Evaluasi komparatif difokuskan pada patologi Cardiomegaly sebagai representasi kelas minoritas yang krusial secara klinis, sebagaimana dirangkum dalam tabel V.

TABEL V
KOMPARASI KINERJA MODEL BERDASARKAN VARIASI FUNGSI KERUGIAN PADA KELAS CARDIOMEGALY

Fungsi Kerugian/Metode	AUROC	Sensitivitas (*Recall*)
Focal Loss ($\alpha = 0{,}25;\ \gamma = 2{,}0$)	0,896	0,824
BCE + *pos_weight* (usulan)	0,901	0,851

Hasil pengujian pada tabel V menunjukkan bahwa implementasi Binary Cross-Entropy (BCE) yang dikombinasikan dengan pembobotan kelas dinamis (pos_weight) serta optimasi ambang batas Youden's J Statistic mencatatkan performa yang lebih superior. Metode usulan mampu mencapai nilai AUROC sebesar 0,901 dan Recall sebesar 0,851, mengungguli pendekatan Focal Loss yang mencatatkan AUROC 0,896 dan Recall 0,824.

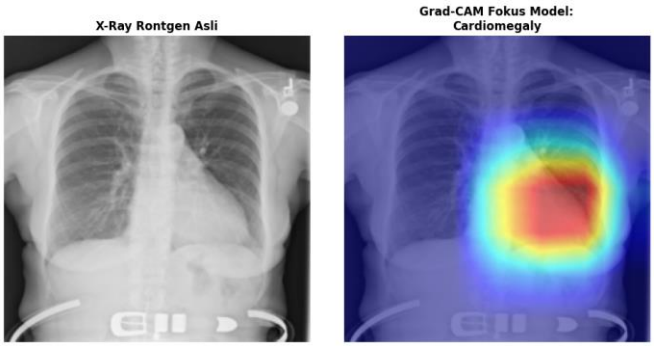
Meskipun Focal Loss menunjukkan performa yang kompetitif (AUROC 0,896 dan sensitivitas 0,824 untuk Cardiomegaly), pendekatan berbasis pembobotan langsung (pos_weight) dalam penelitian ini memiliki beberapa keunggulan strategis: (1) Efisiensi komputasi, karena model tidak perlu mengeksekusi komputasi rambat maju (forward pass) ekstra untuk mengidentifikasi sampel yang sulit dipelajari (hard examples); (2) Transparansi matematis, di mana parameter pos_weight dapat diinterpretasikan secara langsung sebagai rasio ketidakseimbangan kelas untuk proses kalibrasi klinis; dan (3) Stabilitas numerik, mengingat fungsi BCEWithLogitsLoss telah dioptimalkan secara mendalam dalam arsitektur komputasi, sedangkan Focal Loss menuntut penyesuaian hiperparameter tambahan (α dan γ).

Di sisi lain, teknik penyeimbangan data berbasis Generative Adversarial Networks (GAN) atau Synthetic

Minority Over-sampling Technique (SMOTE) tidak dieksplorasi dalam penelitian ini berdasarkan dua pertimbangan utama: (1) Pada set data berskala masif seperti NIH ChestX-ray14, teknik penambahan sampel minoritas rentan memicu fenomena overfitting pada sampel sintetis; dan (2) Dalam konteks radiologi, citra medis sintetis yang dihasilkan oleh model generatif belum tervalidasi sepenuhnya untuk standar keamanan klinis dan berpotensi memperkenalkan artefak visual yang menyesatkan (hallucination). Oleh karena itu, strategi penanganan ketidakseimbangan kelas melalui pembobotan fungsi kerugian terbukti sebagai solusi yang jauh lebih konservatif, efisien, dan aman untuk aplikasi penapisan medis.

H. Analisis Interpretabilitas Klinis

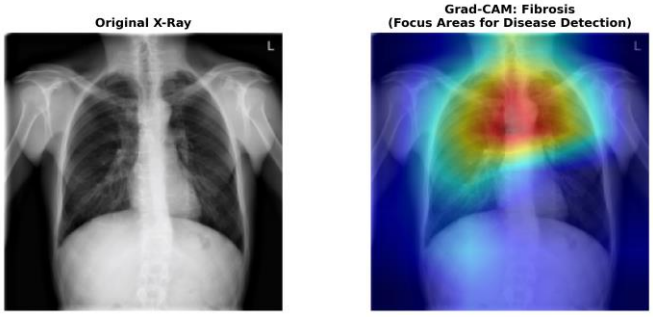
Tantangan utama dalam adopsi model deep learning di ranah medis adalah sifat kotak hitam (black-box) dari jaringan saraf tiruan. Untuk membangun kepercayaan klinis, penelitian ini mengintegrasikan analisis interpretabilitas menggunakan Gradient-weighted Class Activation Mapping (Grad-CAM). Metode ini menghasilkan representasi spasial berupa peta aktivasi (heatmap) yang memvisualisasikan area minat (region of interest/ROI) yang paling memengaruhi keputusan prediktif model pada lapisan konvolusi terakhir.



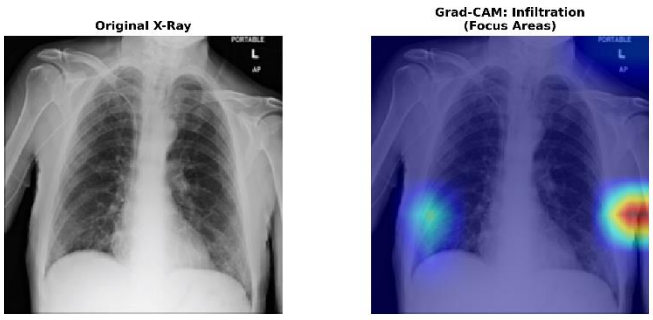
Gambar 5. Visualisasi peta aktivasi Grad-CAM pada pasien dengan diagnosis Kardiomegali (Cardiomegaly). Area berwarna merah hingga kuning menunjukkan area minat (region of interest/ROI) utama yang difokuskan oleh model untuk mengambil keputusan diagnostik.

Berdasarkan visualisasi Grad-CAM pada gambar 5, model menunjukkan kemampuan lokalisasi patologis yang sangat baik meskipun hanya dilatih menggunakan label tingkat citra (image-level labels) tanpa anotasi area spesifik (bounding box). Pada kasus diagnosis Kardiomegali (pembengkakan jantung), peta aktivasi secara presisi terpusat pada area siluet jantung dan batas mediastinum inferior. Hal ini mengonfirmasi bahwa arsitektur MobileNetV2 yang diusulkan tidak sekadar memprediksi berdasarkan artefak pencitraan atau derau (noise) di luar area dada, melainkan secara konsisten mengekstraksi fitur anatomis yang relevan dan sejalan dengan kriteria diagnostik radiologi konvensional.

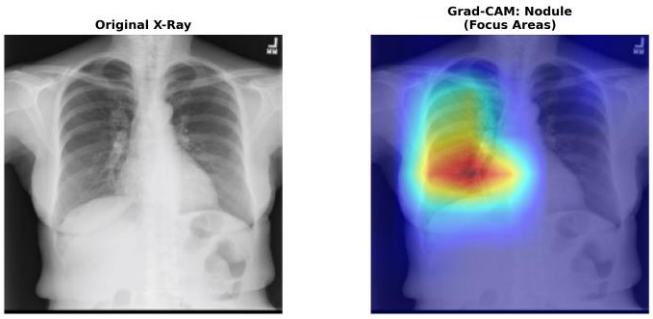
Selain diagnosis Cardiomegaly yang ditunjukkan pada Gambar 5, pengujian lokalisasi Grad-CAM juga dieksekusi untuk empat patologi toraks tambahan yang memiliki signifikansi klinis tinggi: Fibrosis, Infiltration, Nodule, dan Pneumonia. Analisis peta aktivasi (activation maps) ini memvalidasi bahwa model benar-benar mempelajari fitur anatomis yang relevan secara klinis, bukan sekadar mengoptimalkan metrik numerik.



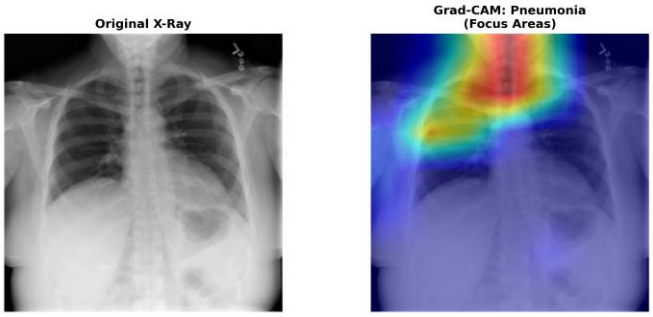
Gambar 6. Visualisasi Grad-CAM untuk Fibrosis. Peta aktivasi menunjukkan fokus model pada area perihilar dan lobus atas-tengah paru-paru dengan pola menyerupai serat (fiber-like pattern). Lokalisasi ini konsisten dengan kriteria temuan radiologis konvensional.



Gambar 7. Visualisasi Grad-CAM untuk *Infiltration*. Model secara akurat memusatkan atensi pada area opasitas fokal. Heterogenitas lokalisasi ini mencerminkan keragaman manifestasi *infiltration* yang dapat muncul di berbagai regio paru-paru.



Gambar 8. Visualisasi Grad-CAM untuk Nodule. Aktivasi model terkonsentrasi secara presisi pada area opasitas bulat kecil (small round opacities). Spesifisitas lokalisasi mikroskopis ini krusial untuk membedakannya dari patologi lain, membuktikan bahwa model mengekstraksi fitur diskriminatif yang tepat.



Gambar 9. Visualisasi Grad-CAM untuk Pneumonia. Area fokus model berada pada titik konsolidasi ruang udara (air-space consolidation) yang tersebar. Distribusi aktivasi ini lebih luas dibandingkan infiltration, merepresentasikan perbedaan karakteristik radiologis antara kedua kondisi patologis tersebut.

Validasi Visual Grad-CAM: Analisis lokalisasi Grad-CAM menunjukkan konsistensi yang presisi dengan kriteria diagnostik radiologi standar. Sebagai representasi, pada kasus Cardiomegaly, peta aktivasi model terpusat murni pada siluet jantung (cardiac silhouette), yang selaras dengan definisi klinis pembesaran jantung dengan rasio kardiotorasik (cardiothoracic ratio) > 0,5. Pada Fibrosis, aktivasi terkonsentrasi pada area perihilar dengan pola serat (reticular pattern), yang selaras dengan manifestasi radiologis fibrosis paru. Kongruensi ini memvalidasi bahwa model secara aktif mengekstraksi representasi dari fitur patologis yang relevan, bukan dari derau (noise) atau artefak pencitraan.

Walaupun demikian, metode interpretabilitas spasial ini memiliki keterbatasan krusial yang perlu diakui: (1) Grad-CAM merepresentasikan area yang memiliki tingkat aktivasi tertinggi, bukan area yang secara kausalistik memicu prediksi model; (2) Pada patologi dengan manifestasi yang tersebar (seperti Infiltration atau Pneumonia), peta aktivasi dapat menyoroti berbagai regio fokal yang memicu ambiguitas interpretasi klinis; dan (3) Metode ini hanya mengkalkulasi gradien pada lapisan konvolusi terakhir, bukan proses transisi keputusan pada lapisan terkoneksi penuh (fully-connected layer). Untuk keperluan peluncuran klinis yang lebih transparan, integrasi metode nilai eksplanasi adaptif (SHAP values) atau mekanisme atensi (attention mechanisms) sangat direkomendasikan pada riset selanjutnya guna memberikan penjelasan prediktif yang tervalidasi secara klinis.

I. Analisis Kalibrasi Probabilitas dan Implikasi Pembobotan

Meskipun penggunaan parameter pos_weight terbukti sukses meningkatkan sensitivitas model, penerapan pembobotan ekstrem pada fungsi kerugian ini secara inheren mendistorsi kalibrasi probabilitas keluaran model (miscalibration). Fenomena ini merupakan kompromi fundamental yang harus dipahami sebelum sistem diimplementasikan secara klinis.

Fungsi kerugian Binary Cross-Entropy yang dibobotkan akan mengalihkan nilai logit menjauh dari distribusi yang konsisten dengan prevalensi kelas sebenarnya. Secara matematis, gradien akan diperkuat secara artifisial untuk kelas minoritas, sehingga model mengompensasinya dengan menghasilkan probabilitas posterior yang lebih tinggi dari realitas klinisnya (overestimation). Sebagai contoh empiris dari pengujian ini, Hernia (dengan prevalensi sebenarnya 0,27%) dapat menghasilkan keluaran probabilitas model di angka 0,40 hingga 0,50. Penyimpangan sistematis ini adalah konsekuensi langsung dari rancangan arsitektur fail-safe yang memprioritaskan sensitivitas.

Secara klinis, hal ini memberikan batasan implementasi yang jelas. Keluaran probabilitas model ini tidak boleh digunakan sebagai skor risiko absolut (absolute risk score) untuk stratifikasi pasien secara individual. Probabilitas 0,45 untuk Hernia tidak bermakna pasien memiliki risiko 45% menderita penyakit tersebut. Akan tetapi, sistem ini sangat valid dan aman jika digunakan untuk perankingan relatif (relative ranking) guna melakukan triase prioritas, serta sebagai sistem pendukung keputusan klinis (clinical decision support) yang menyoroti kasus-kasus berisiko tinggi agar segera ditinjau oleh radiolog.

Untuk riset berkelanjutan yang menargetkan penilaian risiko absolut, penerapan teknik kalibrasi pasca-pelatihan (post-hoc calibration) sangat direkomendasikan. Teknik seperti Platt Scaling (estimasi parameter melalui regresi logistik) atau Temperature Scaling dapat diaplikasikan pada himpunan data kalibrasi terpisah (holdout calibration set). Implementasi kalibrasi ini akan memastikan keluaran probabilitas selaras dengan prevalensi penyakit, sehingga model siap untuk diintegrasikan secara penuh ke dalam alur kerja radiologi.

J. Keterbatasan Penelitian

Meskipun kerangka kerja yang diusulkan mendemonstrasikan kinerja yang menjanjikan, penelitian ini memiliki beberapa keterbatasan yang perlu dipertimbangkan:

1) Ketidakseimbangan Data dan Konsekuensi Penanganannya. Penggunaan parameter pos_weight pada fungsi kerugian Binary Cross-Entropy terbukti sukses meningkatkan sensitivitas pada kelas minoritas, namun menghasilkan kompromi (trade-off) berupa penurunan presisi. Hal ini menunjukkan bahwa pendekatan berbasis pembobotan sangat optimal untuk skenario penapisan medis, namun belum sepenuhnya dapat menggantikan distribusi data yang seimbang.

2) Kalibrasi Probabilitas Keluaran. Intervensi penanganan ketidakseimbangan kelas secara inheren mendisrupsi kalibrasi probabilitas keluaran model (uncalibrated probabilities). Keluaran model cenderung mengalami pergeseran (overestimation) pada kelas minoritas, sehingga belum dapat digunakan secara langsung sebagai metrik penilaian risiko absolut individual tanpa tahapan kalibrasi pasca-pelatihan.

3) Pemodelan Kejadian Bersama (Co-occurrence) yang Implisit. Meskipun kerangka kerja ini dirancang untuk mendeteksi kejadian bersama, pemodelan dilakukan secara implisit melalui probabilitas keluaran sigmoid yang

independen. Arsitektur saat ini belum secara eksplisit memodelkan matriks dependensi antar-label (misalnya menggunakan pendekatan Graph Neural Networks).

4) Generalisasi Lintas Domain. Kemampuan generalisasi model masih terbatas pada distribusi data dari satu sumber institusi (NIH). Model belum diuji pada data eksternal lintas domain untuk memastikan ketangguhannya (robustness) terhadap variasi demografis, perbedaan perangkat radiografi dari berbagai vendor, atau protokol pencitraan di fasilitas kesehatan lain.

5) Ketergantungan pada Anotasi Otomatis. Label patologi dalam set data diekstraksi secara otomatis menggunakan teknik Pemrosesan Bahasa Alami (Natural Language Processing) dari teks laporan radiologi, bukan melalui anotasi manual oleh pakar. Risiko derau label (label noise) tetap berpotensi memengaruhi batas atas performa model.

6) Validasi Klinis Terbatas. Evaluasi diagnostik dalam penelitian ini masih berada pada tahap eksperimental in silico retrospektif. Utilitas klinis dan keamanan sistem belum divalidasi secara prospektif melalui simulasi implementasi pada alur kerja klinis nyata.

7) Implementasi Sistem CAD pada Fasilitas Bersumber Daya Terbatas. Meskipun penelitian ini secara arsitektural telah menekankan efisiensi komputasi MobileNetV2, implementasi operasional sistem CAD di fasilitas kesehatan tingkat pertama tetap menuntut kesiapan infrastruktur tambahan. Hal ini mencakup: (a) ketersediaan peladen atau perangkat tepi (edge devices) untuk inferensi waktu nyata; (b) sistem pengarsipan dan komunikasi citra (PACS) yang terintegrasi penuh dengan alur kerja radiologi; dan (c) jaminan prosedur validasi internal serta kontrol kualitas (quality assurance) yang berkelanjutan. Celah translasi antara pengembangan model komputasional dan kesiapan infrastruktur klinis ini masih perlu dijembatani pada penelitian implementatif di masa mendatang.

K. Rekomendasi Riset Lanjutan

Berdasarkan batasan-batasan di atas, pengembangan penelitian selanjutnya direkomendasikan untuk difokuskan pada area berikut:

- Implementasi Kalibrasi Pasca-Pelatihan: Mengaplikasikan teknik kalibrasi probabilitas seperti Platt Scaling atau Isotonic Regression untuk mengonversi keluaran model menjadi metrik skor risiko absolut yang tepercaya secara klinis.
- Pengembangan Arsitektur Dependensi Label: Mengintegrasikan pemodelan korelasi anatomis patologi secara eksplisit menggunakan Graph Neural Networks guna menangkap probabilitas kejadian bersama dengan lebih komprehensif.
- Uji Validasi Klinis Prospektif dan Simulasi Alur Kerja: Melibatkan pakar radiolog secara aktif dalam studi observasional pada alur kerja klinis nyata. Evaluasi empiris perlu dirancang secara khusus untuk mengukur indikator keberhasilan klinis, yang meliputi: (a) fluktuasi durasi interpretasi citra medis dengan komparasi menggunakan dan tanpa bantuan sistem CAD; (b) rasio penurunan tingkat variabilitas kesepakatan antar-radiolog (inter-observer agreement); (c) dampak sistem terhadap akurasi diagnostik akhir pasien; serta (d) tingkat kepuasan dan persepsi kemudahan penggunaan (user usability) oleh tenaga medis.

V. KESIMPULAN

Penelitian ini berhasil mengembangkan kerangka kerja deep learning multilabel berbasis MobileNetV2 untuk mendeteksi kejadian bersama (co-occurrence) 14 penyakit toraks. Eksperimen menunjukkan bahwa evaluasi kinerja berbasis akurasi global (seperti rata-rata AUROC) sering kali bias terhadap kelas mayoritas dan tidak merepresentasikan keandalan klinis secara utuh. Melalui studi ablasi, terbukti bahwa integrasi pembobotan kelas dinamis dan optimisasi ambang batas berbasis Youden's J Statistic sangat krusial. Meskipun mengorbankan sebagian kecil stabilitas metrik AUROC global, pendekatan ini berhasil menekan angka negatif palsu (false negative) dengan meningkatkan sensitivitas secara substansial pada kelas minoritas berisiko tinggi.

Selain menunjukkan kinerja sensitivitas yang unggul, analisis interpretabilitas menggunakan Grad-CAM mengonfirmasi bahwa kerangka kerja yang diusulkan mampu menunjukkan lokasi area patologis (seperti pembengkakan siluet jantung pada Kardiomegali) dengan presisi anatomis yang tinggi. Keandalan lokalisasi spasial ini, dikombinasikan dengan efisiensi komputasi MobileNetV2 yang ringan (lightweight), menjadikannya instrumen Computer-Aided Diagnosis (CAD) yang potensial dan adaptif untuk diimplementasikan sebagai pembaca kedua (second-reader) di fasilitas kesehatan dengan sumber daya komputasi yang terbatas.

Sebagai arah pengembangan selanjutnya, investigasi lebih lanjut terhadap keseimbangan antara efisiensi dan akurasi (efficiency-accuracy trade-off) sangat direkomendasikan melalui dua rute strategis, dengan mempertimbangkan batasan komputasi pada fasilitas kesehatan: (1) Rute Maksimalisasi Akurasi: Jika target implementasi tidak lagi dibatasi oleh ketersediaan sumber daya, eksplorasi arsitektur berkapasitas tinggi seperti DenseNet121 (95 MB, 102 md/citra) atau Vision Transformer dapat dipertimbangkan guna meningkatkan akurasi diagnostik absolut, dengan pemahaman bahwa hal ini akan meningkatkan kebutuhan infrastruktur secara signifikan; dan (2) Rute Pemertahanan Efisiensi: Jika target implementasi tetap berfokus pada fasilitas kesehatan tingkat pertama dengan sumber daya terbatas, arsitektur MobileNetV2 (71,3 MB, 45 md/citra) sangat direkomendasikan untuk dipertahankan. Namun, pendekatan ini harus diikuti dengan peningkatan resolusi citra masukan menjadi 512 × 512 piksel guna memaksimalkan kemampuan deteksi patologi berukuran mikroskopis seperti nodul dan hernia tanpa mengorbankan efisiensi inferensi

secara substansial.untuk menempuh dua rute pendekatan: (1) Apabila prioritas riset adalah maksimalisasi akurasi absolut, eksplorasi arsitektur berkapasitas tinggi seperti DenseNet121 atau Vision Transformer dapat dipertimbangkan; dan (2) Apabila prioritas riset mempertahankan kelayakan implementasi pada perangkat fasilitas kesehatan primer, penggunaan MobileNetV2 harus dipertahankan, namun disertai dengan peningkatan resolusi citra masukan menjadi 512 × 512 piksel guna mencegah degradasi informasi spasial pada patologi berukuran mikroskopis seperti nodul dan hernia.

DAFTAR PUSTAKA

[1] S. Sajed et al., "The effectiveness of Deep Learning vs. traditional methods for lung disease diagnosis using Chest X-Ray images: A systematic review," Appl. Soft Comput., vol. 147, p. 110817, 2023, doi: 10.1016/j.asoc.2023.110817.

[2] E. Çallı, E. Sogancioglu, B. van Ginneken, K. G. van Leeuwen, and K. Murphy, "Deep Learning for Chest X-Ray analysis: A survey," Med. Image Anal., vol. 72, p. 102125, 2021, doi: 10.1016/j.media.2021.102125.

[3] M. R. Hasan, S. M. A. Ullah, and S. M. R. Islam, "Recent advancement of Deep Learning techniques for pneumonia prediction from Chest X-Ray image," Med. Rep., vol. 7, p. 100106, 2024, doi: 10.1016/j.hmedic.2024.100106.

[4] H. Aljuaid et al., "An experimental comparison of Deep Learning models for pneumonia classification from Chest X-Ray images," Biomed. Signal Process. Control, vol. 112, p. 108742, 2026, doi: 10.1016/j.bspc.2025.108742.

[5] R. Arora et al., "Multi-label classification of Vindr-CXR dataset using transfer-learning approach," Procedia Comput. Sci., vol. 259, pp. 522–531, 2025, doi: 10.1016/j.procs.2025.01.050.

[6] L. A. Gonçalves et al., "DualAttentionNet: A Convolutional Neural Network for thoracic disease classification in Chest X-Rays," Procedia Comput. Sci., vol. 256, pp. 797–804, 2025, doi: 10.1016/j.procs.2025.01.082.

[7] N. Dong, M. Kampffmeyer, H. Su, and E. Xing, "An exploratory study of self-supervised pre-training on partially supervised multi-label classification on Chest X-Ray images," Appl. Soft Comput., vol. 163, p. 111855, 2024, doi: 10.1016/j.asoc.2024.111855.

[8] J. Sun et al., "Multi-label Chest X-Ray image classification based on label co-occurrence," Biomed. Signal Process. Control, vol. 119, p. 109930, 2026, doi: 10.1016/j.bspc.2026.109930.

[9] M. Thapa et al., "An explainable deep-learning-based multi-label image classification for Chest X-Rays," Procedia Comput. Sci., vol. 258, pp. 2425–2434, 2025, doi: 10.1016/j.procs.2025.01.250.

[10] Ş. Öztürk, M. Y. Turalı, and T. Çukur, "HydraViT: Adaptive multi-branch transformer for multi-label disease classification from Chest X-Ray images," Biomed. Signal Process. Control, vol. 100, p. 106959, 2025, doi: 10.1016/j.bspc.2024.106959.

[11] H. Bhatt and M. Shah, "Convolutional Neural Network ensemble model for pneumonia detection using Chest X-Ray images," Healthc. Anal., vol. 3, p. 100176, 2023, doi: 10.1016/j.health.2023.100176.

[12] X. Wang et al., "ChestX-ray8: Hospital-scale Chest X-Ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 2097–2106, doi: 10.1109/CVPR.2017.369.

[13] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 4510–4520, doi: 10.1109/CVPR.2018.00474.

[14] F. M. J. M. Shamrat et al., "High-precision multiclass classification of lung disease through customized MobileNetV2 from Chest X-Ray images," Comput. Biol. Med., vol. 155, p. 106646, 2023, doi: 10.1016/j.compbiomed.2023.106646.

[15] Y. Jin et al., "Deep-learning-based classification of multi-label Chest X-Ray images," Comput. Biol. Med., vol. 157, p. 106683, 2023, doi: 10.1016/j.compbiomed.2023.106683.

[16] G. Chae, J. Lee, and S. B. Kim, "Contrastive learning with hard negative samples for Chest X-Ray classification," Appl. Soft Comput., vol. 165, p. 112101, 2024, doi: 10.1016/j.asoc.2024.112101.

[17] Q. Xu and W. Duan, "DualAttNet: Synergistic fusion of image-level and fine-grained disease attention for multi-label lesion detection in Chest X-Rays," Comput. Biol. Med., vol. 168, p. 107742, 2024, doi: 10.1016/j.compbiomed.2023.107742.