TIN: Terapan Informatika Nusantara
Vol 7, No 1, June 2026, page 228-241
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v7i1.10028

Analisis Sensitivitas Confidence Threshold pada Semi-Supervised FixMatch untuk Klasifikasi Multi-Kelas Citra Chest X-Ray

Ahmad Kurniawan, Muhammad Irsyad, Benny Sukma Negara*, Surya Agustian, Nazruddin Safaat H

Fakultas Sains Dan Teknologi, Prodi Teknik Informatika, Universitas Islam Negeri Sultan Syarif Kasim Riau, Pekanbaru, Indonesia
Email: [1]12250111514@Students.uin-suska.ac.id, [2]irsyadtech@uin-suska.ac.id, [3,*]bsnegara@uin-suska.ac.id,
[4]surya.agustian@uin-suska.ac.id, [5]nazruddin.safaat@uin-suska.ac.id
Email Penulis Korespondensi: bsnegara@uin-suska.ac.id

Abstrak–Optimasi *confidence threshold* dalam *pseudo-labeling* merupakan tantangan teknis kritis pada *Semi-Supervised Learning* (SSL) untuk klasifikasi citra medis multi-kelas, karena threshold yang terlalu ketat membatasi pemanfaatan data tidak berlabel, sementara threshold terlalu longgar memasukkan *pseudo-label* berkualitas rendah ke dalam proses pelatihan. Penelitian ini menerapkan metode FixMatch dengan arsitektur DenseNet-169 sebagai *backbone* untuk mengklasifikasikan tiga kelas penyakit paru COVID-19, Pneumonia, dan Normal, pada kondisi data berlabel yang sangat terbatas. Dataset yang digunakan adalah *Covid19, Pneumonia, and Normal Chest X-Ray Images* dari Mendeley Data sebanyak 5.218 citra dengan pembagian 70% pelatihan, 10% validasi, dan 20% pengujian. Eksperimen dirancang secara sistematis menggunakan tiga proporsi data berlabel (5%, 10%, 15%) dan tiga nilai *confidence threshold* (τ = 0,90; 0,95; 0,99), menghasilkan sembilan skenario percobaan. Hasil menunjukkan bahwa τ = 0,95 dengan 15% data berlabel mencapai performa terbaik (akurasi 97,41%, F1-*Score* 97,49%, AUC 0,9963) karena menyeimbangkan selektivitas *pseudo-label* dengan volume data efektif yang cukup: pada rasio label rendah (5%), volume data berlabel yang terbatas membuat *mask rate* yang lebih rendah di τ = 0,95 tidak cukup dikompensasi sehingga τ = 0,99 unggul tipis, sedangkan pada rasio label tinggi (15%), selektivitas τ = 0,95 menghasilkan *pseudo-label* berkualitas tinggi dengan volume memadai yang mendorong peningkatan generalisasi. Penelitian ini berkontribusi dalam memberikan analisis empiris sensitivitas confidence threshold pada FixMatch untuk klasifikasi multi-kelas CXR dengan data berlabel terbatas. Temuan ini mengungkap bahwa efektivitas *confidence threshold* bersifat kontekstual terhadap ketersediaan label, dan penentuan threshold optimal tidak dapat dilepaskan dari rasio data berlabel yang tersedia.

Kata Kunci: Semi-Supervised Learning; FixMatch; Chest X-Ray; DenseNet-169; Klasifikasi Penyakit Paru

Abstract–Optimizing the confidence threshold in pseudo-labeling is a critical technical challenge in Semi-Supervised Learning (SSL) for multi-class medical image classification. A threshold that is too strict limits the utilization of unlabeled data, whereas a threshold that is too lenient introduces low-quality pseudo-labels into the training process. This study applies the FixMatch method with the DenseNet-169 architecture as the backbone network to classify three lung disease categories COVID-19, Pneumonia, and Normal under conditions of extremely limited labeled data. The dataset used is the *COVID-19, Pneumonia, and Normal Chest X-Ray Images* dataset from Mendeley Data, consisting of 5,218 chest X-ray images, divided into 70% training, 10% validation, and 20% testing sets. The experiments were systematically designed using three labeled-data proportions (5%, 10%, and 15%) and three confidence threshold values (τ = 0.90, 0.95, and 0.99), resulting in nine experimental scenarios. The results demonstrate that τ = 0.95 with 15% labeled data achieved the best performance, obtaining 97.41% accuracy, a 97.49% F1-score, and an AUC of 0.9963. This performance was achieved by balancing pseudo-label selectivity with a sufficient volume of effective training data. At a low labeled-data ratio (5%), the limited amount of labeled data meant that the lower mask rate at τ = 0.95 could not be adequately compensated, allowing τ = 0.99 to perform slightly better. In contrast, at a higher labeled-data ratio (15%), the selectivity of τ = 0.95 produced high-quality pseudo-labels while maintaining sufficient data volume, leading to improved generalization performance. This study contributes an empirical analysis of confidence threshold sensitivity in FixMatch for multi-class chest X-ray classification under limited labeled-data conditions. These findings reveal that the effectiveness of the confidence threshold is highly dependent on the availability of labeled data, and that determining an optimal threshold cannot be separated from the proportion of labeled data available.

Keywords: Semi-Supervised Learning; FixMatch; Chest X-Ray; DenseNet-169; Lung Disease Classification

1. PENDAHULUAN

Radiografi toraks atau *Chest X-Ray* (CXR) merupakan salah satu modalitas pencitraan medis yang paling banyak digunakan dalam evaluasi penyakit paru karena relatif cepat, ekonomis, non-invasif, dan mudah diakses. Pemeriksaan ini menjadi pilihan utama dalam mendeteksi berbagai gangguan paru seperti pneumonia dan COVID-19, yang hingga saat ini masih menjadi masalah kesehatan global. Menurut World Health Organization (2024), pneumonia masih menjadi salah satu penyebab utama kematian akibat penyakit infeksi di dunia, terutama pada kelompok rentan seperti anak-anak dan lansia. Pandemi COVID-19 menambah tekanan luar biasa pada sistem kesehatan, di mana permintaan pemeriksaan CXR melonjak drastis sementara jumlah radiolog tidak bertambah secara proporsional (Li et al., 2023).

Di tengah kondisi ini, model *deep learning* telah menunjukkan performa tinggi dalam mendeteksi pola pada citra CXR dan berpotensi membantu proses diagnosis klinis, termasuk untuk klasifikasi multi-kelas penyakit paru secara simultan (Hmoud et al., 2023; Sahoo et al., 2022). Kondisi ini menunjukkan bahwa pemanfaatan *artificial intelligence* (AI) dan *machine learning* pada analisis citra medis memiliki potensi besar dalam meningkatkan efisiensi, konsistensi, dan kecepatan proses diagnosis klinis, khususnya pada kondisi tingginya beban pemeriksaan radiologi dan keterbatasan tenaga ahli medis (Simon & Aliferis, 2024).

Masalah utama dalam membangun model AI untuk klasifikasi penyakit paru adalah ketimpangan antara jumlah data dan jumlah label. Rumah sakit menyimpan ribuan citra CXR, namun hanya sebagian kecil yang telah dianotasi

TIN: Terapan Informatika Nusantara
Vol 7, No 1, June 2026, page 228-241
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v7i1.10028

oleh radiolog. Keterbatasan data beranotasi disebut sebagai tantangan terbesar dalam pengembangan *deep learning* untuk pencitraan medis (Huang et al., 2023), mengingat proses anotasi memerlukan keahlian khusus dan waktu yang lama. Pelabelan citra medis menjadi hambatan signifikan yang menghambat pengembangan model AI (L. Wang et al., 2021), sehingga pendekatan yang mampu memaksimalkan data tidak berlabel menjadi sangat relevan. Meskipun beberapa Penelitian terkini bahkan membuktikan bahwa model *deep learning* dapat mencapai performa tinggi hanya dengan sekitar 100 sampel berlabel per kelas, kebutuhan anotasi oleh radiolog tetap menjadi hambatan praktis pada banyak institusi kesehatan (Nielsen et al., 2023) Oleh karena itu, pendekatan Semi-Supervised Learning tetap relevan untuk memanfaatkan sejumlah besar data tidak berlabel yang tersedia. Ketika label sangat sedikit, model cenderung *overfitting*, bias terhadap kelas tertentu, dan kehilangan kemampuan generalisasi (Ihler et al., 2024). Konteks klinis menuntut model yang stabil dan andal untuk mendukung keputusan diagnosis.

Beberapa peneliti telah mencoba mengatasi kelangkaan label melalui pendekatan *Semi-Supervised Learning* (SSL). (Sohn et al., 2020) mengusulkan FixMatch, metode yang menggabungkan *pseudo-labeling* dan *consistency regularization* menggunakan pasangan augmentasi lemah dan kuat. FixMatch dilaporkan mencapai akurasi 94,93% pada CIFAR-10 dengan hanya 250 label. Dalam domain medis, (Sahoo et al., 2022) menunjukkan bahwa algoritma COVIDCon berbasis SSL mampu mencapai akurasi 97,07% untuk deteksi COVID-19 dengan label terbatas, membuktikan SSL bisa mengatasi kelangkaan anotasi klinis. (Ihler et al., 2024) menunjukkan bahwa penggunaan FixMatch untuk klasifikasi multi-label pada CXR dapat meningkatkan AUC 1-2% pada setiap kelasnya. Bukti teoritis dan empiris menunjukkan bahwa konsistensi weak-to-strong meningkatkan kualitas *pseudo-label* secara signifikan pada skenario label terbatas (Yang et al., 2023), dan pendekatan SSL berbasis *pseudo-labeling* secara umum terbukti meningkatkan akurasi klasifikasi citra medis hingga 2% dibandingkan metode *baseline* seperti Mean Teacher (K. Liu et al., 2024). Dalam konteks ini, DenseNet-169 dipilih sebagai *backbone* karena unggul pada klasifikasi penyakit paru dibanding arsitektur CNN konvensional (Dalvi et al., 2023), serta lebih stabil pada dataset medis berukuran kecil dibandingkan arsitektur modern seperti ViT dan ConvNeXt yang membutuhkan data pelatihan dalam jumlah besar untuk mencapai performa optimal (Z. Liu et al., 2022; Takahashi et al., 2024).

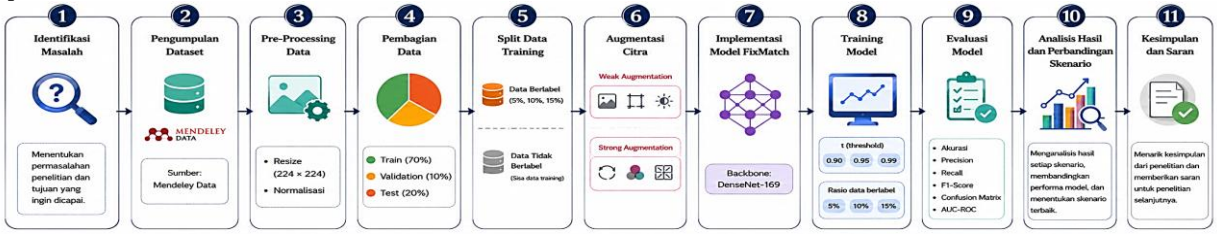
Perkembangan SSL selanjutnya berupaya mengatasi kelemahan utama FixMatch berupa *confidence threshold* statis. FlexMatch (B. Zhang et al., 2021) menyesuaikan threshold secara dinamis per kelas melalui *Curriculum Pseudo Labeling*, sementara FreeMatch (Y. Wang et al., 2023) mengadaptasi threshold secara otomatis tanpa penetapan manual dan terbukti mengungguli FixMatch pada beberapa *benchmark*. Penelitian ini secara sengaja mempertahankan threshold statis karena tujuan utamanya adalah menganalisis sensitivitas nilai threshold terhadap kualitas *pseudo-label* pada klasifikasi multi-kelas CXR secara terkontrol dan sistematis.Meskipun demikian, terdapat celah yang belum terisi. Sebagian besar penelitian SSL pada CXR masih berfokus pada deteksi satu penyakit, sementara klasifikasi multi-kelas secara simultan tetap menjadi tantangan akibat kemiripan pola visual antar kelas (Shamrat et al., 2023). Selain itu, meskipun kedua metode tersebut telah dikembangkan untuk mengatasi kelemahan *threshold* statis FixMatch, kajian sistematis yang memetakan perilaku *threshold* statis pada klasifikasi penyakit paru multi-kelas sebagai dasar perbandingan sebelum beralih ke metode yang lebih adaptif masih belum banyak dilakukan.

Berdasarkan celah tersebut, penelitian ini bertujuan menganalisis sensitivitas *confidence threshold* statis pada FixMatch dengan *backbone* DenseNet-169 untuk klasifikasi tiga kelas penyakit paru (COVID-19, Pneumonia, Normal) pada citra CXR dengan label sangat terbatas. Eksperimen dirancang dengan tiga proporsi data berlabel (5%, 10%, 15%) dan tiga nilai *confidence threshold* (τ = 0,90; 0,95; 0,99), menghasilkan sembilan skenario percobaan. Kontribusi penelitian ini adalah menyediakan kajian empiris mengenai perilaku confidence threshold pada FixMatch untuk klasifikasi multi-kelas Chest X-Ray serta mengungkap pengaruh ketersediaan data berlabel terhadap efektivitas pemanfaatan pseudo-label dalam proses Semi-Supervised Learning.

2. METODOLOGI PENELITIAN

2.1 Kerangka Dasar Penelitian

Penelitian ini dilaksanakan mengikuti alur sistematis yang terdiri dari sebelas tahapan utama yang saling berkaitan. Setiap tahapan dirancang untuk memastikan bahwa proses pengembangan model berlangsung secara terstruktur, dapat direproduksi, dan dapat diverifikasi. Kerangka penelitian pada Gambar 1 menunjukkan tahapan penelitian yang dirancang secara berurutan mulai dari identifikasi masalah hingga evaluasi model dan penarikan kesimpulan. Setiap tahapan saling berkaitan untuk memastikan proses pengembangan model berlangsung secara sistematis dan dapat direproduksi.



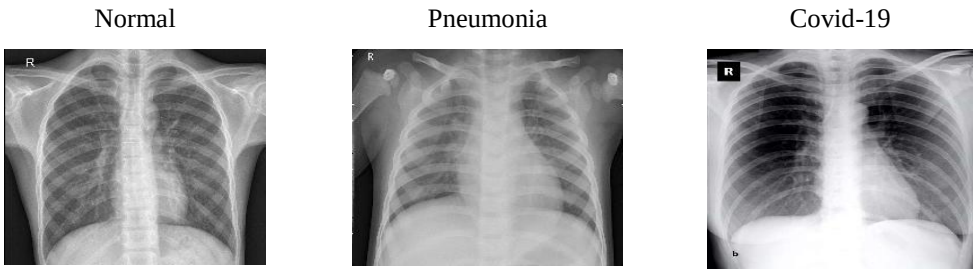
Gambar 1. Kerangka Penelitian

TIN: Terapan Informatika Nusantara
Vol 7, No 1, June 2026, page 228-241
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v7i1.10028

Berdasarkan Gambar 1 penelitian ini dimulai dari identifikasi masalah, yaitu kelangkaan data berlabel pada klasifikasi penyakit paru berbasis CXR. Tahap berikutnya adalah pengumpulan dataset *Covid19, Pneumonia dan Normal Chest X-Ray Images* dari Mendeley Data. Data kemudian masuk ke tahap pra-pemrosesan berupa *resize* 224×224 piksel dan normalisasi nilai piksel. Setelah pra-pemrosesan, dilakukan pembagian data dengan rasio 70% *train*, 10% validasi, dan 20% *test*, di mana data *train* selanjutnya dibagi lagi sesuai proporsi label 5%, 10%, dan 15%. Tahap augmentasi citra menerapkan pasangan *weak* dan *strong augmentation* sesuai mekanisme FixMatch. Model kemudian diimplementasikan menggunakan FixMatch dan DenseNet-169 sebagai *backbone*, lalu dilatih dengan variasi τ (0,90; 0,95; 0,99) dan rasio data berlabel 5%, 10%, 15%. Hasil pelatihan dievaluasi menggunakan akurasi, *precision*, *recall*, F1-*Score*, *confusion matrix*, AUC-ROC, serta perbandingan antar τ. Penelitian ditutup dengan analisis model dan penarikan kesimpulan dan saran.

2.2 Pengumpulan Data

Dataset yang digunakan dalam penelitian ini adalah *Covid19, Pneumonia dan Normal Chest X-Ray Images* yang tersedia secara terbuka melalui platform Mendeley Data (Kumar et al., 2022). Pemilihan dataset ini bukan sekadar pertimbangan ketersediaan, melainkan karena secara langsung mendukung tujuan penelitian yakni mengklasifikasikan tiga kondisi klinis sekaligus dari citra *chest X-ray*: Normal, COVID-19, dan Pneumonia. Total keseluruhan terdiri dari 5.218 citra berformat PNG berukuran 256 × 256 piksel, dengan rincian distribusi yang disajikan pada Gambar 2.



Gambar 2. Dataset (Normal, Pneumonia, COVID-19)

Gambar 2 menampilkan contoh citra dari masing-masing kelas yang digunakan dalam penelitian, yaitu Normal, Pneumonia, dan COVID-19. Ketiga kelas tersebut dipilih karena memiliki karakteristik radiologis yang berbeda sehingga dapat digunakan untuk mengevaluasi kemampuan model dalam membedakan berbagai kondisi paru secara simultan pada skenario klasifikasi multi-kelas. Ada tiga alasan metodologis yang mendasari pemilihan dataset ini. Pertama, statusnya sebagai dataset publik menjamin *reproducibility* peneliti lain dapat mengakses, mereplikasi, maupun membandingkan hasil eksperimen secara langsung. Kedua, citra yang sudah terstruktur per kelas dengan format seragam menyederhanakan penyusunan *pipeline* pelatihan dan meminimalkan risiko kebocoran data antar kelas. Ketiga, kehadiran tiga kelas dalam satu himpunan data menjadi ujian yang relevan bagi model untuk membedakan pola radiologis yang secara visual berdekatan terutama antara Pneumonia dan COVID-19 yang sama-sama menunjukkan opasitas pada paru. Kelayakan akademik dataset ini diperkuat oleh dua studi yang secara resmi direkomendasikan pada halaman Mendeley Data : CheXImageNet (Shastri et al., 2022) yang mengembangkan arsitektur klasifikasi COVID-19 berbasis CNN, dan LiteCovidNet (Kumar et al., 2022) yang merancang model deteksi COVID-19 ringan dari citra X-ray. Kedua studi ini menjadi bukti bahwa dataset tersebut sudah memiliki pijakan akademik yang kokoh dan relevan sebagai basis eksperimen *deep learning* untuk klasifikasi penyakit paru.

2.3 Pra-Pemrosesan Data

Pra-pemrosesan dilakukan untuk menyesuaikan citra dengan kebutuhan input DenseNet-169 sebelum masuk ke tahap pelatihan. Dataset CXR yang digunakan tersedia dalam format PNG dengan resolusi awal 256 × 256 piksel. Seluruh citra kemudian di-*resize* menjadi 224 × 224 piksel menggunakan OpenCV, sesuai spesifikasi input standar DenseNet-169. Penyeragaman ukuran seperti ini diterapkan pada klasifikasi *chest X-ray* berbasis *deep learning* untuk menjaga konsistensi representasi citra antar sampel (Dalvi et al., 2023). Setelah penyesuaian ukuran, citra kemudian dinormalisasi menggunakan statistik ImageNet secara otomatis, yakni *mean* = [0,485; 0,456; 0,406] dan *standard deviation* = [0,229; 0,224; 0,225]. Normalisasi ini diperlukan karena DenseNet-169 diinisialisasi dengan bobot *pretrained* ImageNet, sehingga skala distribusi input perlu selaras dengan kondisi pelatihan awal model (Dalvi et al., 2023). Pendekatan ini diterapkan secara konsisten pada data berlabel maupun data tak berlabel agar pasangan augmentasi *weak* dan *strong* dalam mekanisme FixMatch benar-benar mencerminkan variasi transformasi yang terkontrol. Melalui tahapan pra-pemrosesan tersebut, seluruh citra memiliki ukuran dan distribusi nilai piksel yang seragam sehingga proses ekstraksi fitur oleh DenseNet-169 dapat berlangsung secara lebih stabil dan konsisten pada seluruh sampel.

2.4 Pembagian Data

Data dibagi menggunakan rasio *stratified split* sebesar 70% untuk pelatihan, 10% untuk validasi, dan 20% untuk pengujian guna memastikan distribusi setiap kelas tetap proporsional pada seluruh partisi data. Proses pembagian

TIN: Terapan Informatika Nusantara
Vol 7, No 1, June 2026, page 228-241
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v7i1.10028

dilakukan secara konsisten menggunakan *Random seed* 42 untuk menjaga *reproducibility* penelitian. Tabel 1 merangkum distribusi sampel per kelas dan per partisi.

Tabel 1. Distribusi Dataset

Nama	Total	Train (70%)	Val (10%)	Test (20%)
Covid	1.626	1.138	163	325
Normal	1.802	1.261	180	361
Pneumonia	1.790	1.253	179	358
Total	5.218	3.652	522	1.044

Berdasarkan pada Tabel 1 yang terdiri dari 3.652 sampel pelatihan, kemudian diambil subset berlabel sesuai tiga proporsi yang ditetapkan: 5% (±183 citra), 10% (±365 citra), dan 15% (±548 citra). Sisa data pelatihan diperlakukan sebagai data tak berlabel tanpa informasi kelas yang dimanfaatkan oleh mekanisme *pseudo-labeling* FixMatch. Kombinasi tiga proporsi label dengan tiga nilai *confidence threshold* (τ = 0,90; 0,95; 0,99) menghasilkan sembilan skenario eksperimen yang dijalankan secara sistematis. Penggunaan *stratified split* memastikan setiap kelas terwakili secara proporsional pada seluruh partisi data dan membantu menjaga konsistensi distribusi kelas selama proses pelatihan dan evaluasi model (Osapoetra et al., 2025). Selain itu, pemisahan *train*, *validation*, dan *test set* dilakukan agar evaluasi model menggunakan data yang terpisah dari proses pelatihan sehingga objektivitas evaluasi tetap terjaga (Huang et al., 2023; Nielsen et al., 2023). Untuk mencegah data leakage, seluruh tahapan preprocessing termasuk normalisasi dan augmentasi data diterapkan secara eksklusif berdasarkan statistik yang dihitung hanya dari data pelatihan, kemudian diaplikasikan secara konsisten ke validation set dan test set tanpa pembaruan parameter apapun. Partisi data dilakukan satu kali di awal sebelum proses pelatihan dimulai, sehingga informasi dari validation set maupun test set tidak mempengaruhi keputusan pemodelan dalam bentuk apapun, termasuk pemilihan hyperparameter dan penentuan confidence threshold.

Distribusi ketiga kelas pada dataset ini tergolong near-balanced, dengan selisih jumlah sampel antar kelas tidak melebihi 10% (COVID-19: 1.626, Normal: 1.802, Pneumonia: 1.790). Rasio antara kelas mayoritas dan minoritas hanya sebesar 1,11:1 sehingga belum menunjukkan ketimpangan distribusi yang signifikan. Oleh karena itu, pembobotan kelas (class weighting) pada loss function tidak diterapkan karena distribusi data yang relatif seimbang belum menunjukkan kebutuhan untuk penyesuaian bobot kelas secara khusus (Rajaraman et al., 2022). Sebagai mitigasi tambahan, evaluasi model menggunakan pendekatan macro-averaging pada F1-Score dan AUC-ROC yang memberikan bobot setara pada setiap kelas tanpa dipengaruhi jumlah sampel masing-masing kelas, sehingga performa seluruh kelas tetap terwakili secara proporsional dalam penilaian akhir (Rainio et al., 2024).

2.5 Augmentasi Citra

FixMatch bergantung pada kontras antara augmentasi lemah *weak augmentation* dan augmentasi kuat *strong augmentation*. Kedua jenis augmentasi ini memiliki peran yang berbeda namun saling melengkapi dalam mekanisme pembelajaran konsistensi. Augmentasi lemah dirancang untuk mempertahankan semantik citra: hanya horizontal flip dan translasi ringan (±10% posisi piksel) yang diterapkan. Pseudo-label dihasilkan dari citra dengan augmentasi lemah ini karena representasinya masih dekat dengan citra asli sehingga prediksi model lebih reliabel. Augmentasi kuat menerapkan transformasi yang jauh lebih agresif: horizontal flip, rotasi acak (±15°), perubahan kontras (±10%), random zoom (±30%), translasi lebih besar (±15%), dan penambahan Gaussian noise (standar deviasi 0,1). Model kemudian dilatih untuk menghasilkan prediksi yang konsisten antara keduanya. Strategi ini mengikuti prinsip *weak-to-strong consistency* yang secara teoritis dan empiris terbukti meningkatkan kualitas *pseudo-label* pada skenario label terbatas (Yang et al., 2023). Penggunaan kombinasi augmentasi *weak-strong* ini juga konsisten dengan temuan penelitian terbaru yang menunjukkan bahwa diversitas augmentasi berkontribusi pada generalisasi model yang lebih baik (Alomar et al., 2023). Besaran parameter augmentasi ditetapkan secara konservatif untuk mempertahankan struktur anatomis paru dan menghindari distorsi berlebihan yang dapat mengubah karakteristik klinis citra Chest X-Ray (Yang et al., 2023).

2.6 Implementasi Semi-Supervised Learning FixMatch Dengan DenseNet-169

Pendekatan *Semi-Supervised Learning* (SSL) pada penelitian ini menggunakan metode FixMatch yang menggabungkan *pseudo-labeling* dan *consistency regularization* (Sohn et al., 2020). Formulasi loss function FixMatch diterapkan secara utuh sesuai metode asli tanpa modifikasi pada supervised loss maupun unsupervised loss (Sohn et al., 2020). Keputusan ini dilakukan karena penelitian berfokus pada analisis sensitivitas confidence threshold (τ), sehingga formulasi dasar FixMatch dipertahankan agar perubahan performa dapat diatribusikan secara langsung pada variasi threshold yang diuji. Adaptasi terhadap karakteristik citra Chest X-Ray dilakukan melalui pemilihan DenseNet-169 sebagai backbone serta penggunaan strategi augmentasi yang disesuaikan dengan citra radiologis. Dengan demikian, penyesuaian terhadap domain CXR dilakukan pada desain eksperimen dan representasi fitur, bukan melalui modifikasi loss function (Dalvi et al., 2023; Yang et al., 2023).

Pseudo-label dihasilkan dari citra tidak berlabel menggunakan *weak augmentation*, kemudian digunakan kembali sebagai target supervisi pada citra hasil *strong augmentation*. Penggunaan pasangan *weak* dan *strong augmentation* bertujuan melatih model agar menghasilkan prediksi yang konsisten terhadap variasi transformasi citra *weak-to-strong*

TIN: Terapan Informatika Nusantara
Vol 7, No 1, June 2026, page 228-241
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v7i1.10028

consistency (Yang et al., 2023). Pseudo-label hanya digunakan apabila probabilitas prediksi tertinggi melebihi *confidence threshold* τ sehingga kualitas pseudo-label tetap terjaga (W. Zhang et al., 2022). Penerapan FixMatch pada klasifikasi citra *Chest X-Ray* terbukti mampu menghasilkan F1-score COVID-19 sebesar 0,94 hanya dengan 80 sampel berlabel per kelas, menunjukkan bahwa mekanisme *confidence threshold* berkontribusi signifikan terhadap kualitas pseudo-label pada data berlabel terbatas (Sajun et al., 2022).

1. Mekanisme inti FixMatch pada data tidak berlabel dinyatakan sebagai berikut:

$$\frac{1}{B_u}\sum_{b=1}^{B_u} 1\left(max\left(\mathcal{P}_m\left(\mathcal{Y}|\alpha(B_u)\right)\right) \geq \tau\right) H\left(\arg max\left(\mathcal{P}_m\left(\mathcal{Y}|\alpha(B_u)\right)\right), \mathcal{P}_m\left(\mathcal{Y}|A(B_u)\right)\right) \quad (1)$$

   Pada rumus ini, $B_u$merupakan batch data tidak berlabel. Fungsi indikator $\mathbf{1}(\cdot)$bernilai 1 jika kondisi terpenuhi dan 0 jika tidak. $\mathcal{P}_m(\mathcal{Y}\mid\cdot)$merepresentasikan distribusi probabilitas prediksi model terhadap kelas $\mathcal{Y}$, di mana $\alpha(B_u)$adalah citra tidak berlabel setelah weak augmentation dan $A(B_u)$adalah citra tidak berlabel setelah strong augmentation. τ adalah confidence threshold, yaitu ambang batas keyakinan minimum agar pseudo-label dapat diterima. $H(\cdot,\cdot)$merupakan cross-entropy loss antara dua distribusi, sedangkan $\arg\max(\cdot)$menunjukkan kelas dengan probabilitas prediksi tertinggi yang dijadikan pseudo-label.
2. Pseudo-label pada *weak augmentation* diperoleh menggunakan persamaan berikut:

$$\hat{\mathcal{Y}}b = \arg max\left(\mathcal{P}_w\left(\mathcal{Y}|\alpha(B_u)\right)\right) \quad (2)$$

   $\hat{\mathcal{Y}}b$ adalah pseudo-label yang dihasilkan untuk sampel ke-b. $\mathcal{P}_w(\mathcal{Y}\mid\alpha(B_u))$merupakan probabilitas prediksi model pada citra hasil weak augmentation, di mana $\alpha(B_u)$adalah citra tidak berlabel yang telah melalui transformasi weak augmentation. Pseudo-label ditentukan dengan mengambil kelas yang memiliki nilai probabilitas tertinggi dari distribusi prediksi tersebut.
3. Prediksi model pada citra hasil *strong augmentation* dinyatakan sebagai berikut:

$$\mathcal{P}_s\left(\mathcal{Y}|A(B_u)\right) \quad (3)$$

   Rumus tersebut merupakan probabilitas prediksi model pada citra yang telah melalui strong augmentation. $A(B_u)$merepresentasikan citra tidak berlabel setelah diterapkan serangkaian transformasi agresif berupa rotasi, perubahan kontras, zoom, translasi, dan penambahan Gaussian noise
4. Supervised loss pada data berlabel dihitung menggunakan *cross-entropy* sebagai berikut:

$$\mathcal{L}_s = \frac{1}{B_l}\sum_{b=1}^{B_l} H\left(\mathcal{Y}_{B_l}, \mathcal{P}_w(\mathcal{Y}|\mathcal{X}_i)\right) \quad (4)$$

   $\mathcal{L}_s$ adalah supervised loss yang dihitung dari data berlabel. $B_l$merupakan batch data berlabel, sedangkan $H(\cdot,\cdot)$adalah cross-entropy loss yang mengukur kesalahan antara prediksi model dan label asli. $\mathcal{Y}_{B_l}$merupakan label asli atau ground truth dari data berlabel, sementara $\mathcal{P}_w(\mathcal{Y}\mid\mathcal{X}_i)$adalah probabilitas prediksi model pada citra berlabel $\mathcal{X}_i$, di mana $\mathcal{X}_i$merupakan citra berlabel ke-i.
5. Unsupervised loss pada data tidak berlabel dihitung menggunakan pseudo-label dan confidence masking:

$$\mathcal{L}_u = \frac{1}{B_u}\sum_{b=1}^{B_u} 1(max(\mathcal{P}_w) \geq \tau)\, H\left(\hat{\mathcal{Y}}b, \mathcal{P}_s\left(\mathcal{Y}|A(B_u)\right)\right) \quad (5)$$

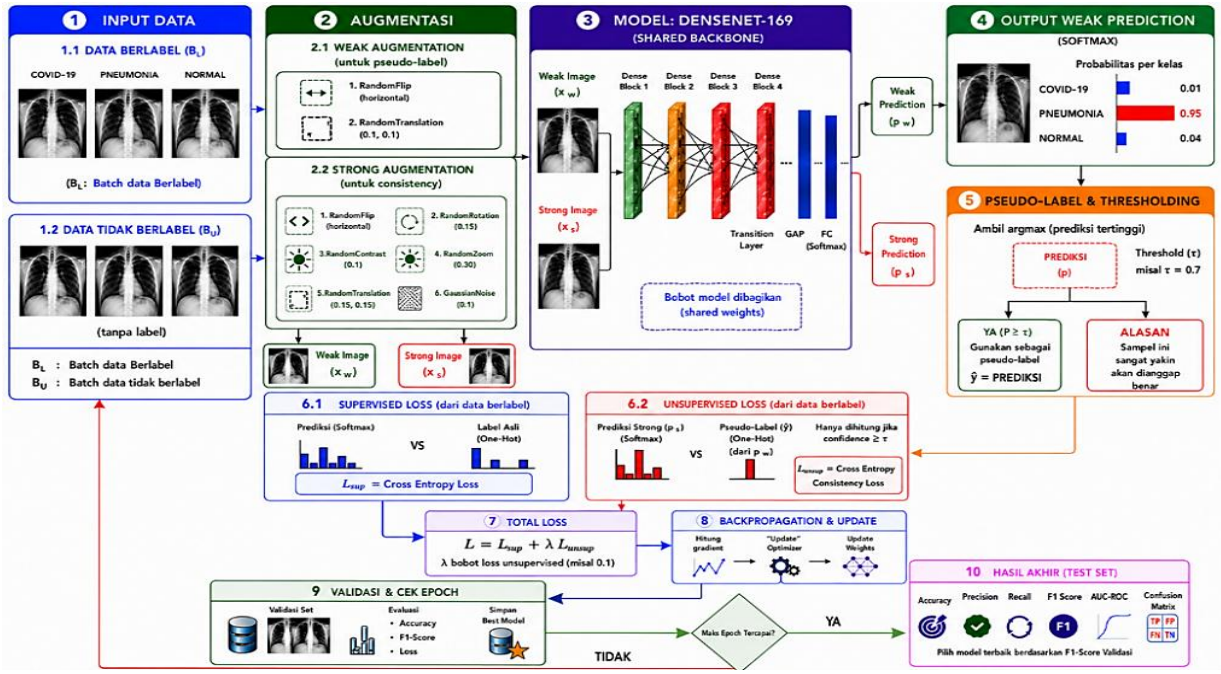
   $\mathcal{L}_u$ adalah unsupervised loss yang dihitung dari data tidak berlabel. $B_u$merupakan batch data tidak berlabel. $1(\max(\mathcal{P}_w) \geq \tau)$adalah fungsi masking yang memastikan pseudo-label hanya digunakan apabila probabilitas prediksi tertinggi memenuhi atau melampaui confidence threshold τ. $\mathcal{P}_w$adalah probabilitas prediksi model pada citra hasil weak augmentation, $\hat{\mathcal{Y}}_b$merupakan pseudo-label untuk sampel ke-b, dan $\mathcal{P}_s(\mathcal{Y}\mid A(B_u))$adalah probabilitas prediksi model pada citra hasil strong augmentation.
6. Total loss pada metode FixMatch diperoleh dari penjumlahan *supervised loss* dan *unsupervised loss* yang dikontrol menggunakan koefisien λ_u (Sohn et al., 2020):

$$\mathcal{L} = \mathcal{L}_s + \lambda_u \cdot \mathcal{L}_u \quad (6)$$

   $\mathcal{L}$ adalah total loss keseluruhan yang dioptimalkan selama proses pelatihan. $\mathcal{L}_s$merupakan supervised loss yang berasal dari data berlabel, sedangkan $\mathcal{L}_u$merupakan unsupervised loss yang berasal dari data tidak berlabel. $\lambda_u$adalah koefisien pengali yang mengontrol seberapa besar kontribusi unsupervised loss terhadap total loss, dengan nilai yang digunakan dalam penelitian ini sebesar 1,0.

Implementasi pada penelitian ini menggunakan DenseNet-169 sebagai *backbone classifier* dengan bobot *pretrained ImageNet*. DenseNet-169 dipilih karena memiliki kemampuan representasi fitur yang baik pada klasifikasi citra medis dan mampu mempertahankan aliran informasi antar lapisan melalui mekanisme *dense connectivity* (Dalvi et al., 2023). Proses pelatihan dilakukan selama 100 epoch menggunakan *optimizer* Adam dengan *learning rate* awal 1×10^{-4}, *batch size* data berlabel 16, rasio *unlabeled* ($\mu = 7$), serta variasi *confidence threshold* ($\tau = 0{,}90; 0{,}95; 0{,}99$). Alur implementasi semi-supervised learning FixMatch dengan DenseNet-169 ditunjukkan pada Gambar 3.

TIN: Terapan Informatika Nusantara
Vol 7, No 1, June 2026, page 228-241
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v7i1.10028



Gambar 3. Alur kerja FixMatch dengan DenseNet-169

Berdasarkan Gambar 3, proses pelatihan FixMatch memanfaatkan data berlabel dan data tidak berlabel secara bersamaan. Data tidak berlabel menghasilkan pseudo-label melalui *weak augmentation*, kemudian digunakan sebagai target pada *strong augmentation* untuk menghitung *unsupervised loss* yang digabungkan dengan *supervised loss* dalam proses optimasi model.

2.7 Training Model Dengan Variasi Parameter

Pelatihan dilakukan menggunakan kombinasi tiga *confidence threshold* (τ = 0,90; 0,95; 0,99) dan tiga proporsi data berlabel (5%, 10%, 15%) sehingga menghasilkan sembilan skenario eksperimen. Nilai threshold tinggi dipilih karena FixMatch memanfaatkan *pseudo-label* dengan tingkat keyakinan tinggi untuk menjaga kualitas supervisi pada data tidak berlabel (Sohn et al., 2020). Setiap skenario dimulai dengan inisialisasi ulang model dan *optimizer* menggunakan *random seed* 42 untuk menjaga *reproducibility* dan mencegah kebocoran informasi antar skenario. Parameter pelatihan yang digunakan secara konsisten pada seluruh eksperimen dirangkum pada Tabel 2.

Tabel 2. Konfigurasi Hyperparameter

Hyperparameter	Nilai
Confidence Threshold (τ)	0,90 / 0,95 / 0,99
Unsupervised Loss Weight (λu)	1,0
Rasio Unlabeled (μ)	7
Batch Size Labeled	16
Batch Size Unlabeled	112
Optimizer	Adam
Learning Rate	1×10^{-4}
Maks Epoch	100

Berdasarkan Tabel 2 nilai rasio unlabeled μ = 7 ditetapkan mengikuti konfigurasi default pada penelitian FixMatch asli (Sohn et al., 2020) dan dipertahankan secara konsisten pada seluruh skenario eksperimen. Dengan batch size data berlabel sebesar 16, nilai tersebut menghasilkan 112 sampel tidak berlabel pada setiap iterasi pelatihan sehingga proporsi pemanfaatan data berlabel dan data tidak berlabel tetap sejalan dengan konfigurasi yang telah tervalidasi pada penelitian FixMatch sebelumnya (Sohn et al., 2020). Karena fokus utama penelitian ini adalah menganalisis sensitivitas *confidence threshold (τ)*, parameter μ dipertahankan konstan pada seluruh eksperimen agar perubahan performa yang diamati dapat dikaitkan secara langsung dengan variasi threshold yang diuji. Pendekatan serupa juga digunakan pada berbagai penelitian lanjutan berbasis FixMatch yang mempertahankan konfigurasi inti metode untuk mengevaluasi pengaruh parameter tertentu secara terisolasi (Ihler et al., 2024).

Sebagai pembanding terhadap pendekatan Semi-Supervised Learning, penelitian ini juga mengimplementasikan baseline supervised-only menggunakan arsitektur DenseNet-169 yang sama. Pada skenario baseline, model dilatih hanya menggunakan data berlabel tanpa memanfaatkan data tidak berlabel, pseudo-labeling, maupun komponen unsupervised loss yang terdapat pada FixMatch. Seluruh konfigurasi pelatihan, termasuk optimizer Adam, learning rate 1×10^{-4}, batch size, jumlah epoch, pembagian data, dan random seed dipertahankan identik dengan eksperimen

TIN: Terapan Informatika Nusantara
Vol 7, No 1, June 2026, page 228-241
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v7i1.10028

FixMatch agar perbedaan performa yang diperoleh dapat diatribusikan secara langsung pada pemanfaatan data tidak berlabel melalui mekanisme Semi-Supervised Learning.

2.8 Evaluasi dan Analisis Model

Evaluasi akhir dilakukan menggunakan *test set* yang sepenuhnya terpisah dari proses pelatihan dan validasi agar hasil evaluasi bersifat objektif. Seluruh skenario diuji pada *test set* yang sama untuk menjaga konsistensi perbandingan antar konfigurasi. Penelitian ini menggunakan pendekatan evaluasi multi-metrik karena satu metrik tunggal belum cukup merepresentasikan performa model pada kondisi data berlabel terbatas (Rainio et al., 2024). Evaluasi pertama menggunakan akurasi *accuracy*, *precision*, *recall*, dan *F1-Score* makro dengan pendekatan *macro averaging* sehingga setiap kelas memiliki bobot yang sama.

1. Akurasi dihitung menggunakan persamaan berikut:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

   Persamaan tersebut menunjukkan proporsi prediksi benar terhadap seluruh data pengujian.
2. *Precision* digunakan untuk mengukur ketepatan prediksi model terhadap suatu kelas:

$$Precision = \frac{TP}{TP + FP} \quad (8)$$

3. *Recall* digunakan untuk mengukur kemampuan model dalam mendeteksi seluruh sampel positif:

$$Recall = \frac{TP}{TP + FN} \quad (9)$$

4. *F1-Score* digunakan sebagai rata-rata antara *precision* dan *recall*:

$$F1\text{-}Score = \frac{2\ x\ Precision\ x\ Recall}{Precision + Recall} \quad (10)$$

5. *confusion matrix* digunakan untuk menganalisis pola kesalahan klasifikasi antar kelas, khususnya antara Pneumonia dan COVID-19 yang memiliki karakteristik radiologis serupa. Visualisasi *confusion matrix* ditampilkan dalam bentuk *heatmap*, di mana warna biru muda menunjukkan jumlah prediksi rendah, sedangkan warna biru tua menunjukkan jumlah prediksi yang lebih tinggi. Semakin gelap warna pada sel diagonal utama, semakin baik performa klasifikasi model karena menunjukkan jumlah prediksi benar yang lebih besar.
6. *AUC-ROC macro-average* dengan pendekatan *one-vs-rest* untuk mengukur kemampuan diskriminatif probabilistik model pada berbagai ambang keputusan. Nilai AUC berada pada rentang 0–1, di mana nilai yang semakin mendekati 1 menunjukkan kemampuan klasifikasi yang semakin baik.
7. *Grad-CAM* (*Gradient-weighted Class Activation Mapping*) pada model terbaik di setiap *threshold* untuk menelusuri area citra yang paling berkontribusi terhadap prediksi model berdasarkan gradien pada lapisan konvolusi terakhir DenseNet-169 (Selvaraju et al., 2020). Pada visualisasi *Grad-CAM*, warna merah dan merah tua menunjukkan area dengan aktivasi tertinggi yang paling diperhatikan model, warna kuning dan hijau menunjukkan aktivasi sedang, sedangkan warna biru menunjukkan area dengan kontribusi paling rendah terhadap proses prediksi (Umair et al., 2021).

3. HASIL DAN PEMBAHASAN

3.1 Hasil Keseluruhan Eksperimen

Pengujian dilakukan pada sembilan skenario eksperimen yang merupakan kombinasi tiga proporsi data berlabel (5%, 10%, 15%) dengan tiga nilai *confidence threshold* (τ = 0,90; 0,95; 0,99). Seluruh skenario dievaluasi menggunakan test set yang identik agar perbandingan antar konfigurasi bersifat objektif. Rekapitulasi performa model disajikan pada Tabel 3.

Tabel 3. Ringkasan Hasil Sembilan Skenario Eksperimen pada Data Uji

Threshold (τ)	Label	Citra Berlabel	Best Val F1	Akurasi	F1-Macro	Avg Mask Rate
0,90	5%	182	95,43	93,49	93,56	87,7
0,90	10%	365	98,13	96,74	96,83	91,8
0,90	15%	548	98,51	96,65	96,73	93,3
0,95	5%	182	95,87	94,73	94,85	85,7
0,95	10%	365	97,94	96,26	96,37	91,7
0,95	15%	548	98,13	97,41	97,49	89,6
0,99	5%	182	95,86	95,02	95,11	66,2
0,99	10%	365	97,95	96,74	96,83	85,4
0,99	15%	548	97,94	97,22	97,30	83,5

TIN: Terapan Informatika Nusantara
Vol 7, No 1, June 2026, page 228-241
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v7i1.10028

Berdasarkan Tabel 3, seluruh skenario berhasil melampaui akurasi 93% meskipun hanya menggunakan 182–548 citra berlabel dari 3.652 sampel pelatihan. Hasil terbaik dicapai oleh skenario τ = 0,95 dengan rasio 15% (akurasi 97,41% dan F1-Macro 97,49%), sedangkan performa terendah tercatat pada τ = 0,90 rasio 5% (akurasi 93,49%). Meskipun selisih antara skenario terbaik dan terburuk hanya 3,93 poin persentase, peningkatan proporsi data berlabel secara konsisten berkontribusi positif terhadap performa model di seluruh nilai threshold. Peningkatan paling signifikan terjadi pada transisi dari 5% ke 10% data berlabel, sedangkan transisi dari 10% ke 15% menunjukkan perlambatan yang mengindikasikan model mulai mendekati titik saturasi pada rentang label tersebut. Pola ini berlaku konsisten di ketiga nilai threshold, menunjukkan bahwa pengaruh rasio data berlabel lebih dominan dibandingkan pengaruh pemilihan nilai threshold itu sendiri.

Tabel 4. Perbandingan SSL FixMatch vs Supervised

Rasio Label	Citra Berlabel	Supervised Acc (%)	FixMatch Acc (%) [τ]	Gain Acc (PP)	Supervised F1 (%)	FixMatch F1 (%) [τ]	Gain F1 (PP)
5%	182	95,40	95,02 [τ 99]	-0,38	95,53	95,11 [τ 99]	-0,42
10%	365	96,36	96,74 [τ 90]	+0,38	96,46	96,83 [τ 90]	+0,37
15%	548	97,13	97,41 [τ 95]	+0,28	97,21	97,49 [τ 95]	+0,28

Berdasarkan Tabel 4, efektivitas FixMatch bersifat kontekstual terhadap jumlah data berlabel yang tersedia. Pada rasio 5%, FixMatch (τ = 0,99) menghasilkan akurasi 95,02% dan F1-Macro 95,11%, lebih rendah 0,38–0,42 poin persentase dibanding *supervised-only* (95,40% dan 95,53%). Penurunan ini mencerminkan kondisi *cold start* yang dikenal dalam literatur SSL: ketika label terlalu sedikit, model belum memiliki representasi fitur yang cukup stabil untuk menghasilkan *pseudo-label* yang andal, sehingga sinyal *unsupervised* justru menambahkan *noise* ke dalam pelatihan (Yang et al., 2023). Temuan ini menunjukkan bahwa SSL tidak selalu memberikan keuntungan pada rasio label yang sangat rendah. Sebaliknya, pada rasio 10% dan 15%, FixMatch secara konsisten mengungguli supervised-only dengan gain akurasi masing-masing +0,38 pp dan +0,28 pp. Hasil ini menunjukkan bahwa pemanfaatan data tidak berlabel mulai memberikan manfaat ketika jumlah data berlabel cukup untuk membangun representasi fitur yang stabil. Dengan hanya 10–15% data berlabel, FixMatch mampu menyamai bahkan melampaui pendekatan supervised-only tanpa menambah kebutuhan anotasi.

3.2 Perbandingan Antar Confidence Threshold

Analisis perbandingan antar nilai confidence threshold dilakukan untuk mengungkap pola interaksi antara sensitivitas τ dan ketersediaan data berlabel yang menjadi pertanyaan utama penelitian ini. Pola interaksi ini tidak bersifat linear dan bervariasi bergantung pada jumlah citra berlabel yang tersedia pada setiap skenario.

Tabel 5. Perbandingan F1-Macro Berdasarkan τ dan Rasio Label

Threshold (τ)	Rasio 5%	Rasio 10%	Rasio 15%
0,90	93,56	96,83	96,73
0,95	94,85	96,37	97,49
0,99	95,11	96,83	97,30

Berdasarkan Tabel 5 menunjukkan bahwa pengaruh nilai τ terhadap performa model bersifat tidak linear dan dipengaruhi oleh jumlah data berlabel. Pada rasio 5%, F1-Macro meningkat seiring kenaikan τ, yaitu dari 93,56% (τ = 0,90), 94,85% (τ = 0,95), hingga 95,11% (τ = 0,99). Hal ini dipengaruhi oleh mask rate awal yang semakin kecil pada threshold tinggi, yaitu 36,5% (τ = 0,90), 11,0% (τ = 0,95), dan 3,5% (τ = 0,99). Dengan jumlah label yang sangat terbatas, pseudo-label yang lebih selektif pada τ = 0,99 membantu menjaga kualitas prediksi dibanding pseudo-label yang lebih banyak tetapi berpotensi noisy pada τ = 0,90. Kondisi ini diperparah oleh mask rate awal yang sangat rendah pada τ = 0,99 (3,5% di epoch pertama), yang berarti hampir seluruh data tidak berlabel tidak diikutsertakan dalam pelatihan awal. Meskipun pseudo-label yang masuk berkualitas tinggi, volume sinyal unsupervised yang sangat terbatas membuat manfaatnya baru terasa setelah model memiliki representasi fitur yang cukup matang di epoch-epoch berikutnya.

Pada rasio 10%, τ = 0,90 dan τ = 0,99 menghasilkan F1-Macro yang sama, yaitu 96,83%, sedangkan τ = 0,95 sedikit lebih rendah (96,37%). Hal ini menunjukkan bahwa dengan 365 citra berlabel, model mulai stabil sehingga pengaruh threshold tidak terlalu signifikan. Pada rasio 15%, τ = 0,95 memberikan hasil terbaik dengan F1-Macro 97,49%, lebih tinggi dibanding τ = 0,99 (97,30%) dan τ = 0,90 (96,73%). Dengan jumlah label yang lebih banyak, threshold moderat mampu memberikan keseimbangan antara kualitas pseudo-label dan pemanfaatan data tidak berlabel. Secara keseluruhan, pengaruh confidence threshold bergantung pada jumlah data berlabel yang tersedia. Threshold tinggi cenderung lebih efektif pada rasio label rendah, sedangkan threshold moderat memberikan hasil terbaik ketika jumlah data berlabel meningkat. Temuan ini menunjukkan bahwa tidak terdapat satu nilai threshold yang optimal pada seluruh skenario, sehingga pemilihannya perlu disesuaikan dengan ketersediaan data berlabel.

TIN: Terapan Informatika Nusantara
Vol 7, No 1, June 2026, page 228-241
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v7i1.10028

3.3 Analisis Per-Kelas

Analisis per kelas memberikan pemahaman yang lebih mendalam tentang kekuatan dan kelemahan model dalam membedakan masing-masing kondisi klinis. Analisis ini dilakukan dengan membandingkan dua skenario ekstrem: skenario tertinggi (τ = 0,95, rasio label 15%, 548 citra berlabel) dan skenario terendah (τ = 0,90, rasio label 5%, 182 citra berlabel). Perbandingan antara kedua skenario ini memberikan gambaran tentang dampak konfigurasi terhadap setiap kelas secara individual. Hasil analisis per kelas disajikan pada Tabel 6.

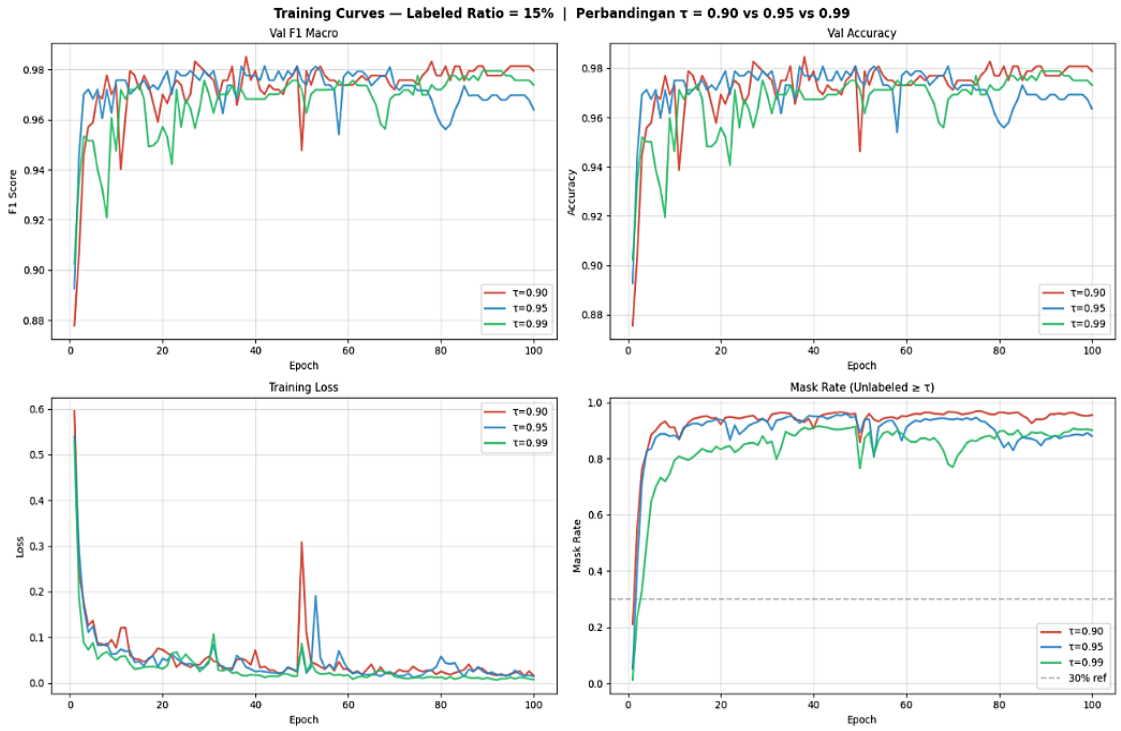
Tabel 6. Analisis Per-Kelas: τ=0,95 Rasio 15% vs τ=0,90 Rasio 5%

Kelas	Precision 0,95 | 15%	Recall 0,95 | 15%	F1 0,95 | 15%	Precision 0,90 | 5%	Recall 0,90 | 5%	F1 0,90 | 5%
COVID-19	1,00	1,00	1,00	0,94	1,00	0,97
Normal	0,96	0,96	0,96	0,92	0,91	0,91
Pneumonia	0,96	0,96	0,96	0,95	0,91	0,93
Macro Avg	0,97	0,97	0,97	0,94	0,94	0,94

Berdasarkan Tabel 6, Kelas COVID-19 konsisten paling mudah dideteksi di seluruh skenario mencapai F1 sempurna 1,00 pada skenario terbaik dan recall 1,00 bahkan pada skenario terburuk. Sebaliknya, kesalahan klasifikasi terbanyak terjadi pada pasangan Normal dan Pneumonia akibat kemiripan pola visual keduanya. Pada skenario terburuk, recall kedua kelas ini turun ke 0,91; sementara pada skenario terbaik naik ke 0,96 dengan hanya 29 kesalahan dari 1.044 citra uji (2,78%). Seluruh kesalahan klasifikasi pada skenario terbaik terjadi pada pasangan Normal dan Pneumonia, sedangkan kelas COVID-19 terklasifikasi sempurna. Hal ini mengonfirmasi bahwa peningkatan proporsi data berlabel memberikan dampak paling besar pada kemampuan model membedakan Normal dan Pneumonia, bukan pada deteksi COVID-19 yang sudah kuat bahkan pada kondisi label sangat terbatas.

3.4 Kurva Training dan Dinamika Maks Rate

Analisis dinamika pelatihan dilakukan untuk memahami bagaimana model belajar dari waktu ke waktu dan bagaimana mekanisme *pseudo-labeling* bekerja sepanjang proses pelatihan. Pengamatan ini penting karena memberikan wawasan tentang stabilitas konvergensi dan efektivitas pemanfaatan data tidak berlabel. Kurva pelatihan yang menampilkan training loss, validation F1, dan mask rate per rasio label disajikan pada Gambar 4.



Gambar 4. Kurva Training Loss, Validation F1, dan Mask Rate per Rasio Label

Berdasarkan Gambar 4 pada Kurva pelatihan menunjukkan karakteristik konvergensi yang berbeda antar konfigurasi. Pada skenario 5% label, seluruh threshold memulai dengan validation F1 yang rendah di epoch pertama τ = 0,90 tercatat 0,287, τ = 0,95 tercatat 0,815, dan τ = 0,99 tercatat 0,905. Meskipun τ = 0,99 memulai dengan val F1 awal tertinggi, mask rate epoch pertamanya hanya 3,5%, jauh di bawah τ = 0,90 yang mencapai 36,5%. Kondisi ini mengungkap trade-off yang terjadi dalam mekanisme pseudo-labeling: mask rate yang rendah pada τ = 0,99 menjaga kualitas pseudo-label namun membatasi volume pemanfaatan data tidak berlabel, sehingga model lebih bergantung

TIN: Terapan Informatika Nusantara
Vol 7, No 1, June 2026, page 228-241
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v7i1.10028

pada sinyal supervised yang jumlahnya juga terbatas. Sebaliknya, mask rate tinggi pada τ = 0,90 meningkatkan volume pemanfaatan data tidak berlabel, namun juga meningkatkan kemungkinan masuknya pseudo-label yang kurang selektif sehingga performa model menjadi lebih sensitif terhadap kualitas pseudo-label. Pada skenario 10% dan 15% label, seluruh threshold menunjukkan konvergensi yang lebih stabil. Best validation F1 yang dicapai meningkat signifikan: τ = 0,90 dengan 15% label mencapai best val F1 98,51%, τ = 0,95 dengan 15% label 98,13%, dan τ = 0,99 dengan 15% label 97,94%. Namun hasil ini menunjukkan bahwa, meskipun τ = 0,90 mencapai best val F1 tertinggi (98,51%), performa test set-nya (96,73%) lebih rendah dari τ = 0,95 (97,49%). Hasil pelatihan menunjukkan bahwa threshold yang terlalu rendah menghasilkan jumlah pseudo-label yang lebih banyak, namun kualitas prediksinya cenderung kurang stabil. Kondisi ini menunjukkan bahwa peningkatan jumlah pseudo-label pada threshold rendah tidak selalu menghasilkan performa terbaik pada data uji. Meskipun nilai validasi relatif tinggi, kemampuan generalisasi model masih dipengaruhi oleh kualitas pseudo-label yang digunakan selama proses pelatihan. Temuan ini memperkuat pentingnya evaluasi akhir menggunakan test set yang sepenuhnya terpisah dari proses pelatihan.

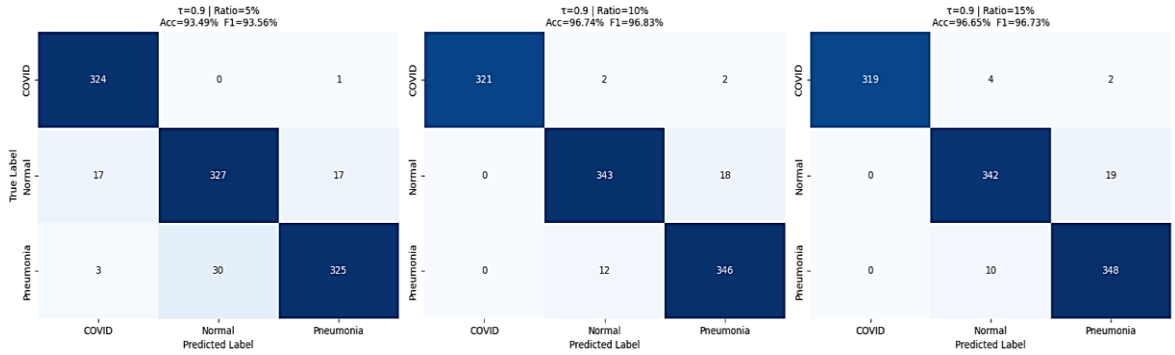
Tabel 7. Rata-rata Mask Rate Selama Pelatihan

Rasio Label	Data Citra	τ = 0,90	τ = 0,95	τ = 0,99
5%	182	87,7	85,7	66,2
10%	365	91,8	91,7	85,4
15%	548	93,3	89,6	83,5

Tabel 7 mengonfirmasi bahwa semakin tinggi τ, semakin rendah rata-rata mask rate. Penurunan paling dramatis terjadi pada rasio 5% dari 87,7% (τ = 0,90) turun ke 66,2% (τ = 0,99), selisih 21,5 poin. Pada rasio 15%, selisih antara τ = 0,90 dan τ = 0,99 menyempit menjadi 9,8 poin (93,3% vs 83,5%), karena label yang lebih banyak membantu model membangun probabilitas prediksi yang lebih stabil pada tahap awal pelatihan.

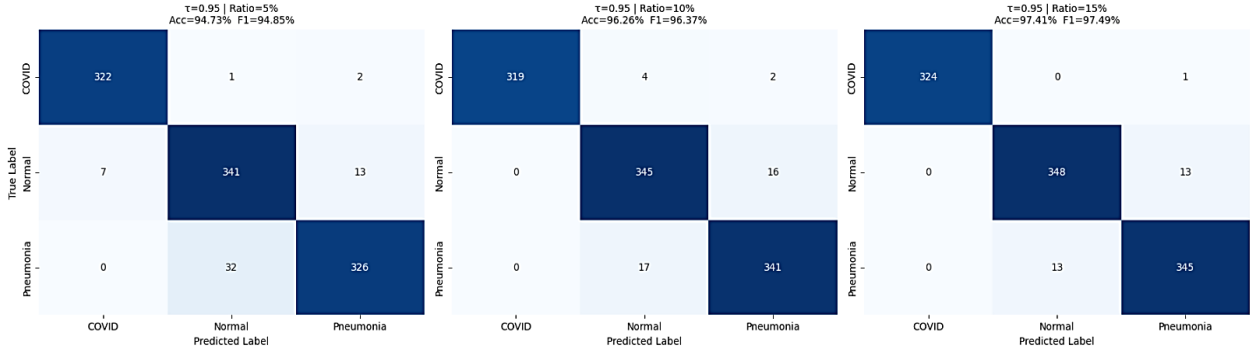
3.5 Confusion Matrix dan AUC-ROC

Visualisasi *confusion matrix* dan kurva AUC-ROC memberikan perspektif yang lebih komprehensif tentang pola kesalahan model dan kemampuan diskriminatifnya. Kedua evaluasi ini saling melengkapi, *confusion matrix* menunjukkan pola kesalahan spesifik antar kelas, sementara AUC-ROC menilai kemampuan diskriminatif probabilistik secara keseluruhan. *Confusion matrix* untuk setiap konfigurasi threshold dan rasio label disajikan pada Gambar 5, 6, dan 7.



Gambar 5. Confusion Matrix threshold 90 per Rasio Label

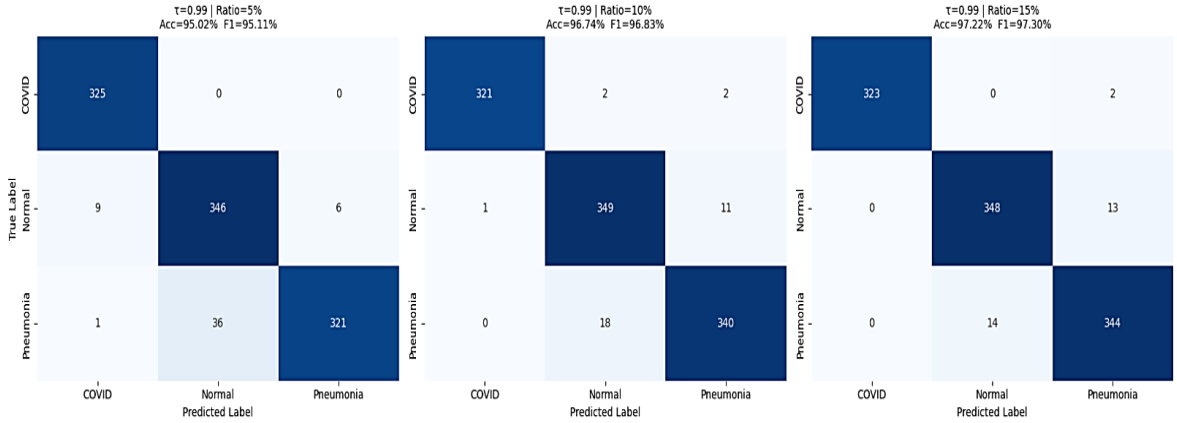
Berdasarkan Gambar 5, threshold τ = 0,90 menunjukkan kesalahan klasifikasi terkonsentrasi pada pasangan Normal-Pneumonia. Pada rasio 5%, Normal hanya mencapai 90,9% dan Pneumonia 90,8%, sementara COVID-19 sempurna (325/325). Peningkatan rasio label ke 10% dan 15% secara konsisten mengurangi kesalahan pada pasangan tersebut.



Gambar 6. Confusion Matrix threshold 95 per Rasio Label

TIN: Terapan Informatika Nusantara
Vol 7, No 1, June 2026, page 228-241
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v7i1.10028

Berdasarkan Gambar 6, threshold τ = 0,95 menghasilkan performa terbaik pada rasio 15% dengan COVID-19 sempurna (325/325), Normal 95,8% (346/361), dan Pneumonia 96,1% (344/358). Threshold moderat ini mampu menyeimbangkan kualitas pseudo-label dan volume pemanfaatan data tidak berlabel.



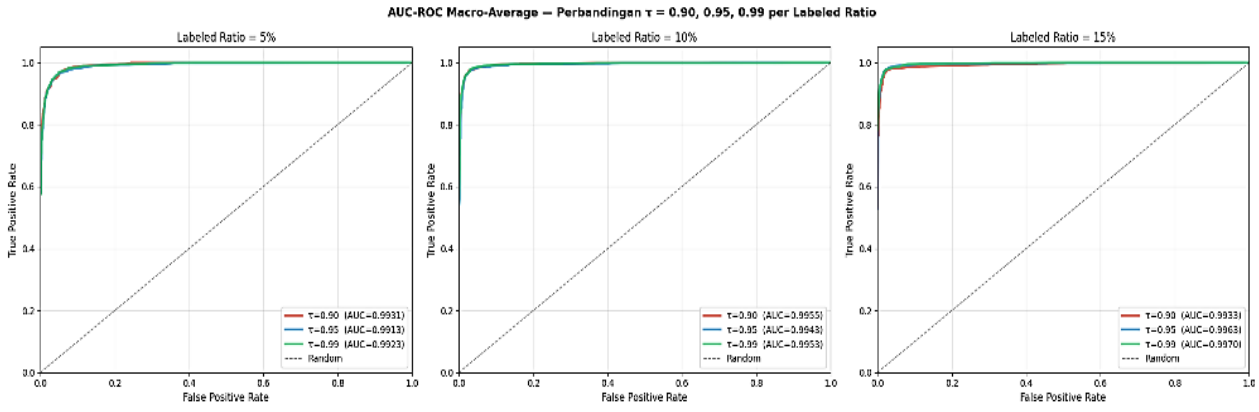
Gambar 7. Confusion Matrix threshold 99 per Rasio Label

Confusion matrix seluruh skenario mengonfirmasi pola yang konsisten: kesalahan klasifikasi terkonsentrasi pada pasangan Normal–Pneumonia, sementara kelas COVID-19 hampir selalu terklasifikasi dengan benar di seluruh konfigurasi threshold maupun rasio label. Hal ini mengonfirmasi bahwa penambahan data berlabel paling berdampak pada diskriminasi Normal dan Pneumonia, bukan pada deteksi COVID-19 yang sudah kuat sejak label sangat terbatas. Selain *confusion matrix*, evaluasi AUC-ROC memberikan perspektif tambahan tentang kemampuan diskriminatif probabilistik model. Nilai AUC macro-average untuk seluruh skenario dirangkum pada Tabel 8, sedangkan kurva AUC-ROC secara visual disajikan pada Gambar 8.

Tabel 8. Nilai AUC Macro-Average Seluruh Skenario

Rasio Label	Data Citra	AUC (τ = 0,90)	AUC (τ = 0,95)	AUC (τ = 0,99)
5%	182	0,9931	0,9913	0,9923
10%	365	0,9955	0,9943	0,9953
15%	548	0,9933	0,9963	0,9970

Berdasarkan Tabel 8, seluruh skenario menghasilkan AUC ≥ 0,9913 yang menunjukkan kemampuan diskriminatif model sangat tinggi di semua konfigurasi. Nilai AUC tertinggi dicapai oleh τ = 0,99 rasio 15% (0,9970), diikuti τ = 0,95 rasio 15% (0,9963). Meskipun τ = 0,99 menghasilkan AUC tertinggi, τ = 0,95 dipilih sebagai konfigurasi terbaik karena menghasilkan akurasi dan F1-Macro tertinggi pada data uji. Secara umum, peningkatan rasio label berkontribusi positif terhadap nilai AUC di seluruh threshold.



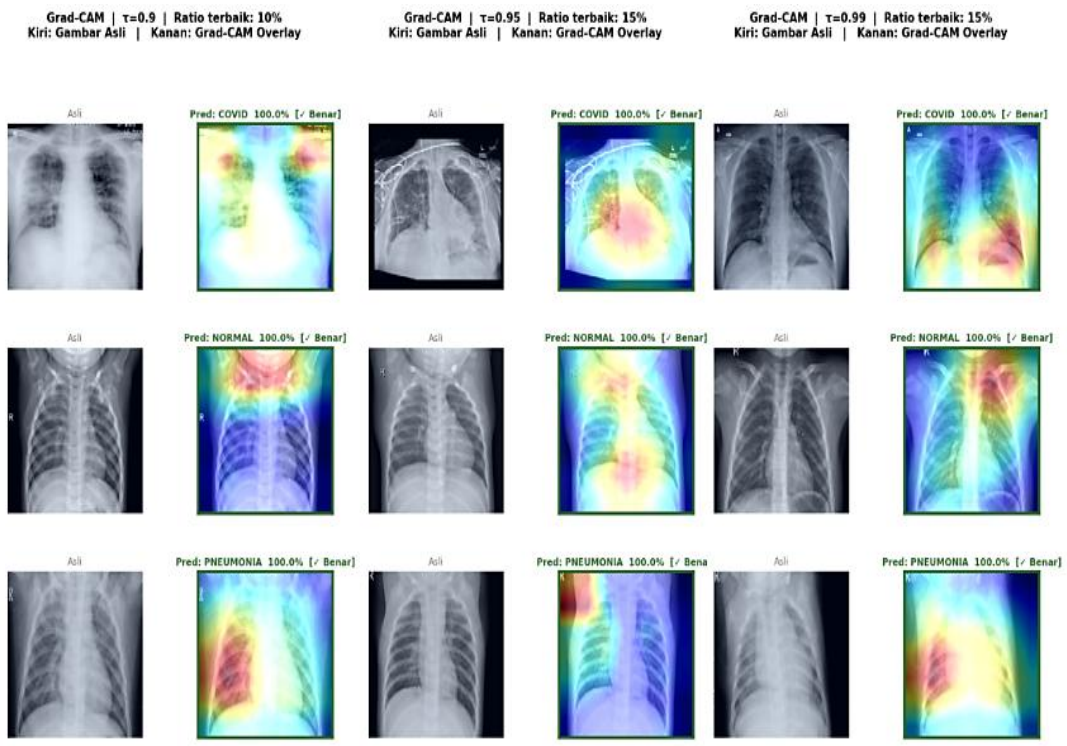
Gambar 8. Kurva AUC-ROC Macro-Average per Rasio Label

Berdasarkan Gambar 8, seluruh kurva AUC-ROC mendekati sudut kiri atas yang mengindikasikan kemampuan diskriminatif probabilistik yang tinggi pada semua skenario. Secara keseluruhan, Tabel 8 dan Gambar 8 mengonfirmasi bahwa meskipun τ = 0,99 rasio 15% menghasilkan AUC tertinggi (0,9970), F1-Score *test set*-nya (97,30%) tetap lebih rendah dibanding τ = 0,95 (97,49%), menunjukkan bahwa performa probabilistik tidak selalu sejalan dengan hasil klasifikasi akhir. Bahkan pada skenario label paling sedikit (5%), AUC minimum 0,9913 membuktikan bahwa kombinasi *pretrained* DenseNet-169 dengan mekanisme FixMatch mampu membangun representasi fitur yang kuat meski supervisi sangat terbatas.

TIN: Terapan Informatika Nusantara
Vol 7, No 1, June 2026, page 228-241
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v7i1.10028

3.6 Visualisasi Grad-CAM

Untuk memahami mekanisme keputusan model secara lebih mendalam, dilakukan analisis Grad-CAM (Gradient-weighted Class Activation Mapping) yang memberikan visualisasi tentang area citra yang paling berkontribusi terhadap prediksi model. Analisis ini penting untuk memvalidasi secara komputasional bahwa model berfokus pada region yang relevan secara anatomis, bukan pada artefak atau noise. Grad-CAM diterapkan pada lapisan konvolusi terakhir DenseNet-169. Model yang digunakan untuk analisis Grad-CAM dipilih berdasarkan performa terbaik setiap threshold: $\tau = 0{,}90$ menggunakan model rasio 10% (best val F1 98,13%), $\tau = 0{,}95$ dan $\tau = 0{,}99$ masing-masing menggunakan model rasio 15% (best val F1 98,13% dan 97,94%). Hasil visualisasi Grad-CAM ditampilkan secara representatif pada Gambar 9.



Gambar 9. Visualisasi Grad-CAM pada Model Terbaik Setiap Threshold

Berdasarkan analisis Gambar 9, secara umum aktivasi model terkonsentrasi pada area toraks dan paru yang sesuai dengan area utama pada citra *Chest X-Ray*. Pada kelas COVID-19, aktivasi model terlihat tersebar pada area paru yang secara visual berkaitan dengan pola opasitas difus yang umum ditemukan pada kasus COVID-19. Pada kelas Pneumonia, aktivasi lebih terlokalisasi pada area konsolidasi fokal yang secara visual menyerupai area konsolidasi pada kasus pneumonia. Pada kelas Normal, distribusi aktivasi lebih merata tanpa fokus patologis spesifik, mencerminkan bahwa model mengenali ketidakhadiran pola abnormal sebagai fitur tersendiri.

Pola aktivasi yang relevan secara anatomis ini mengindikasikan bahwa model mempelajari pola visual yang konsisten dengan area anatomis paru melalui pemanfaatan data tidak berlabel pada mekanisme *pseudo-labeling* FixMatch. Meskipun demikian, interpretasi ini bersifat komputasional dan belum divalidasi secara klinis oleh radiolog (Ihongbe et al., 2024; Selvaraju et al., 2020). Validasi klinis dengan melibatkan radiolog berpengalaman merupakan langkah lanjutan yang penting sebelum sistem ini diterapkan dalam konteks diagnosis nyata. Selain itu, perlu diperhatikan bahwa visualisasi Grad-CAM hanya merepresentasikan area aktivasi berdasarkan gradien pada lapisan konvolusi terakhir dan bukan penjelasan kausal atas keputusan model secara menyeluruh. Interpretasi yang berhati-hati terhadap keterbatasan ini diperlukan agar hasil Grad-CAM tidak diartikan secara berlebihan sebagai bukti yang konklusif mengenai mekanisme keputusan model.

3.7 Pembahasan

Hasil penelitian ini membuktikan bahwa kombinasi FixMatch dengan DenseNet-169 efektif untuk klasifikasi multi-kelas penyakit paru berbasis CXR dalam kondisi label yang sangat terbatas. Perbandingan langsung terhadap baseline *supervised-only* pada Tabel 4 menunjukkan bahwa FixMatch mulai mengungguli *supervised-only* pada rasio 10% dan 15% dengan gain hingga +0,38 pp, meskipun pada rasio 5% masih sedikit di bawah akibat kondisi *cold start* di mana sinyal *unsupervised* baru memberikan manfaat ketika representasi fitur dasar sudah cukup terbentuk (Yang et al., 2023).

Perbandingan dengan penelitian SSL sebelumnya juga memperkuat temuan ini. Sahoo et al. (2022) mencapai akurasi 97,07% untuk deteksi *single-class* COVID-19, sedangkan penelitian ini mencapai 97,41% pada klasifikasi *multi-kelas* tiga kondisi sekaligus yang secara tugas jauh lebih kompleks. Sajun et al. (2022) melaporkan F1-Score

TIN: Terapan Informatika Nusantara
Vol 7, No 1, June 2026, page 228-241
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v7i1.10028

COVID-19 sebesar 0,94 dengan 80 label per kelas menggunakan FixMatch, sementara penelitian ini mencapai F1-Score sempurna (1,00) pada skenario terbaik, menunjukkan kontribusi positif pemilihan DenseNet-169 dan strategi augmentasi yang disesuaikan untuk citra radiologis.

Dibandingkan Kumar et al. (2022) yang melaporkan akurasi 98,82% menggunakan LiteCovidNet dengan seluruh data berlabel, penelitian ini lebih rendah 1,41 poin persentase. Namun penelitian ini hanya menggunakan 15% data berlabel (548 dari 3.652 citra), membuktikan bahwa FixMatch mampu mendekati performa *supervised learning* penuh dengan kebutuhan anotasi yang jauh lebih sedikit, yang secara praktis lebih relevan terhadap kondisi klinis nyata di mana data berlabel sulit diperoleh (Huang et al., 2023).

4. KESIMPULAN

Penelitian ini menunjukkan bahwa efektivitas confidence threshold pada metode FixMatch tidak bersifat universal, melainkan bergantung pada ketersediaan data berlabel yang digunakan selama pelatihan. Hasil eksperimen memperlihatkan bahwa peningkatan jumlah data berlabel memberikan pengaruh yang lebih dominan terhadap performa dibandingkan perubahan nilai threshold itu sendiri. Pada kondisi label yang sangat terbatas, threshold yang lebih tinggi cenderung lebih efektif karena menghasilkan pseudo-label yang lebih selektif, sedangkan ketika jumlah data berlabel meningkat, threshold moderat mampu memberikan keseimbangan yang lebih baik antara kualitas pseudo-label dan pemanfaatan data tidak berlabel. Temuan ini menunjukkan bahwa pemilihan threshold optimal perlu mempertimbangkan rasio data berlabel dan tidak dapat ditetapkan secara terpisah dari kondisi ketersediaan label. Analisis per-kelas menunjukkan bahwa model mampu membedakan kelas COVID-19 secara konsisten pada seluruh skenario, sementara kesalahan klasifikasi masih didominasi oleh pasangan kelas Normal dan Pneumonia yang memiliki karakteristik radiologis serupa. Hasil tersebut mengindikasikan bahwa tantangan utama pada klasifikasi multi-kelas citra Chest X-Ray bukan hanya keterbatasan jumlah label, tetapi juga tingginya kemiripan visual antar kelas penyakit tertentu. Penelitian ini memiliki keterbatasan karena eksperimen hanya dilakukan pada satu dataset publik dan satu arsitektur backbone, yaitu DenseNet-169, sehingga kemampuan generalisasi model terhadap data dari rumah sakit atau populasi yang berbeda belum dapat dipastikan secara langsung. Oleh karena itu, penelitian selanjutnya perlu melakukan validasi eksternal menggunakan dataset multi-institusi serta membandingkan FixMatch dengan pendekatan threshold adaptif seperti FlexMatch dan FreeMatch untuk memperoleh pemahaman yang lebih komprehensif mengenai penerapan Semi-Supervised Learning pada klasifikasi citra medis. Secara akademis, penelitian ini berkontribusi dalam menunjukkan bahwa efektivitas confidence threshold pada FixMatch dipengaruhi oleh ketersediaan data berlabel, sehingga pemilihan threshold optimal perlu mempertimbangkan kondisi label yang tersedia selama proses pelatihan.

REFERENCES

Alomar, K., Aysel, H. I., & Cai, X. (2023). Data Augmentation in Classification and Segmentation : A Survey and New Strategies. *Journal of Imaging*, *9*(2), 46. https://doi.org/10.3390/jimaging9020046

Dalvi, P. P., Edla, D. R., & Purushothama, B. R. (2023). Diagnosis of Coronavirus Disease From Chest X-Ray Images Using DenseNet-169 Architecture. *SN Computer Science*, *4*(3), 1–6. https://doi.org/10.1007/s42979-022-01627-7

Hmoud, M., Sheikh, A., Dandan, O. Al, Sami, A., Shamayleh, A., Jalab, H. A., & Ibrahim, R. W. (2023). Multi - class deep learning architecture for classifying lung diseases from chest X - Ray and CT images. *Scientific Reports*, 1–14. https://doi.org/10.1038/s41598-023-46147-3

Huang, S. C., Pareek, A., Jensen, M., Lungren, M. P., Yeung, S., & Chaudhari, A. S. (2023). Self-supervised learning for medical image classification: a systematic review and implementation guidelines. *Npj Digital Medicine*, *6*(1). https://doi.org/10.1038/s41746-023-00811-0

Ihler, S., Kuhnke, F., Kuhlgatz, T., & Seel, T. (2024). Distribution-Aware Multi-Label FixMatch for Semi-Supervised Learning on CheXpert. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2295–2304. https://doi.org/10.1109/CVPRW63382.2024.00235

Ihongbe, I. E., Shereen, F., Mahmoud, T. F., & Rajasekaran, A. (2024). Evaluating Explainable Artificial Intelligence (XAI) techniques in chest radiology imaging through a human-centered Lens. *PLoS ONE*, *19*, 1–27. https://doi.org/10.1371/journal.pone.0308758

Kumar, S., Shastri, S., Kansal, I., Singh, K., Popli, R., & Mansotra, V. (2022). LiteCovidNet : A lightweight deep neural network model for detection of COVID-19 using X-ray images. *International Journal of Imaging Systems and Technology*, *32*(February), 1464–1480. https://doi.org/10.1002/ima.22770

Li, M., Jiang, Y., Zhang, Y., & Zhu, H. (2023). Medical image analysis using deep learning algorithms. *Frontiers in Public Health*, *11*(November), 1–28. https://doi.org/10.3389/fpubh.2023.1273253

Liu, K., Liu, J., & Liu, S. (2024). Enhanced Semi-Supervised Medical Image Classification Based on Dynamic Sample Reweighting and Pseudo-Label Guided Contrastive Learning (DSRPGC). *Mathematics*, *12*(22), 3572. https://doi.org/10.3390/math12223572

Liu, Z., Mao, H., Christoph, C. W., Trevor, F., Saining, D., & Berkeley, U. C. (2022). A ConvNet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11976–11986. https://doi.org/10.1109/CVPR52688.2022.01167

Nielsen, M., Wenderoth, L., Sentker, T., & Werner, R. (2023). Self-Supervision for Medical Image Classification :

TIN: Terapan Informatika Nusantara
Vol 7, No 1, June 2026, page 228-241
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v7i1.10028

State-of-the-Art Performance with ~ 100 Labeled Training Samples per Class. *Bioengineering*, *10*(8), 895. https://doi.org/10.3390/bioengineering10080895

Osapoetra, L. O., Moslemi, A., Moore-palhares, D., Halstead, S., Alberico, D., Hwang, A., Sannachi, L., & Curpen, B. (2025). *End-to-end CNN-based deep learning enhances breast lesion characterization using quantitative ultrasound (QUS) spectral parametric images*. 1–13.

Rainio, O., Jarmo, T., & Klen, R. (2024). Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, 1–14. https://doi.org/10.1038/s41598-024-56706-x

Rajaraman, S., Ganesan, P., & Antani, S. (2022). Deep learning model calibration for improving performance in class-imbalanced medical image classification tasks. *PLoS ONE*, *17(1): e02*, 1–23. https://doi.org/10.1371/journal.pone.0262838

Sahoo, P., Roy, I., Ahlawat, R., Irtiza, S., & Khan, L. (2022). Potential diagnosis of COVID - 19 from chest X - ray and CT findings using semi - supervised learning. *Physical and Engineering Sciences in Medicine*, *45*(1), 31–42. https://doi.org/10.1007/s13246-021-01075-2

Sajun, A. R., Zualkernan, I., & Sankalpa, D. (2022). Investigating the Performance of FixMatch for COVID-19 Detection in Chest X-rays. *Applied Sciences (Switzerland)*, *12*(9). https://doi.org/10.3390/app12094694

Selvaraju, R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *International Journal of Computer Vision*, *128*(2), 336–359. https://doi.org/10.1007/s11263-019-01228-7

Shamrat, M., Javed, F. M., Azam, S., Karim, A., Ahmed, K., & Bui, F. M. (2023). High-precision multiclass classification of lung disease through customized MobileNetV2 from chest X-ray images. *Computers in Biology and Medicine*, *155*(February), 106646. https://doi.org/10.1016/j.compbiomed.2023.106646

Shastri, S., Kansal, I., Kumar, S., Singh, K., Popli, R., & Mansotra, V. (2022). CheXImageNet : a novel architecture for accurate classification of Covid - 19 with chest x - ray digital images using deep convolutional neural networks. *Health and Technology*, 193–204. https://doi.org/10.1007/s12553-021-00630-x

Simon, G. J., & Aliferis, C. (2024). *Artificial Intelligence and Machine Learning in Health Care and Medical Sciences*. Springer.

Sohn, K., Berthelot, D., Li, Chun, L., Zhang, Z., Carlini, N., Cubuk, E. D., Kurakin, A., Zhang, H. H., & Raffel, C. (2020). Semi-Supervised : FixMatch. *Advances in Neural Information Processing Systems*, *37*(10), 1575–1585. https://arxiv.org/abs/2001.07685

Takahashi, S., Sakaguchi, Y., Kouno, N., Takasawa, K., Ishizu, K., & Akagi, Y. (2024). Comparison of Vision Transformers and Convolutional Neural Networks in Medical Image Analysis : A Systematic Review. *Journal of Medical Systems*, *48*(1), 1–22. https://doi.org/10.1007/s10916-024-02105-8

Umair, M., Khan, M. S., Ahmed, F., Baothman, F., Alqahtani, F., Alian, M., & Ahmad, J. (2021). *Detection of COVID-19 Using Transfer Learning and Grad-CAM Visualization on Indigenously Collected*. *June*. https://doi.org/https://doi.org/10.3390/s21175813

Wang, L., Guo, D., Wang, G., & Zhang, S. (2021). Annotation-Efficient Learning for Medical Image Segmentation based on Noisy Pseudo Labels and Adversarial Learning. *IEEE Transactions on Medical Imaging*, *40*(8), 2235–2246. https://doi.org/10.1109/TMI.2020.3047807

Wang, Y., Chen, H., Heng, Q., Hou, W., Fan, Y., Savvides, M., & Shinozaki, T. (2023). FreeMatch: Self-adaptive thresholding in semi-supervised learning. *International Conference on Learning Representations*, 1–20.

Yang, L., Feng, L., Shi, Y., Qi, L., & Zhang, W. (2023). Revisiting Weak-to-Strong Consistency. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7236–7246. https://doi.org/10.1109/CVPR52729.2023.00699

Zhang, B., Yang, W., Hou, W., Wu, H., Wang, J., Okumura, M., & Shinozaki, T. (2021). FlexMatch : Boosting Semi-Supervised Learning with Curriculum Pseudo Labeling. *Advances in Neural Information Processing Systems*, *34*(NeurIPS), 18408–18419.

Zhang, W., Zhu, L., Hallinan, J., Makmur, A., Zhang, S., Cai, Q., & Ooi, Chin, B. (2022). BoostMIS : Boosting Medical Image Semi-supervised Learning with Adaptive Pseudo Labeling and Informative Active Annotation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 20666–20676. https://doi.org/10.1109/CVPR52729.2022.02007

World Health Organization. (2024). Pneumonia. Retrieved from https://www.who.int/health-topics/pneumonia