



Application of Semi-Supervised Deep Learning with AdaMatch for Classification of Lung Disease on Chest X-ray Image

Penerapan Semi-Supervised Deep Learning dengan AdaMatch untuk Klasifikasi Penyakit Paru-paru pada Citra X-ray Dada

**Hilya Zalwana¹, Benny Sukma Negara^{2*}, Muhammad Irsyad³,
Suwanto Sanjaya⁴, Muhammad Fikry⁵**

^{1,2,3,4}Program Studi Teknik Informatika, Fakultas Sains dan Teknologi,
Universitas Islam Negeri Sultan Syarif Kasim Riau, Indonesia

E-Mail: ¹12250125451@students.uin-suska.ac.id, ²bsnegara@uin-suska.ac.id,
³irsyadtech@uin-suska.ac.id, ⁴suwantosanjaya@uin-suska.ac.id, ⁵muhammad.fikry@uin-suska.ac.id

Received May 09th 2026; Revised Jun 10th 2026; Accepted Jun 21th 2026; Available Online Jun 26th 2026

Corresponding Author: Benny Sukma Negara

Copyright © 2026 by Authors, Published by Institut Riset dan Publikasi Indonesia (IRPI)

Abstract

Lung diseases, such as Coronavirus Disease 2019 (COVID-19) and pneumonia, remain major health challenges that require rapid and accurate diagnosis. Chest X-ray (CXR) imaging is widely used for pulmonary examination; however, its interpretation still depends heavily on radiologists and the availability of labeled medical data. The limited availability of labeled data poses a significant challenge for developing deep learning models based on supervised learning. This study proposes a semi-supervised deep learning approach using AdaMatch with DenseNet-169 for multiclass classification of CXR images into COVID-19, Pneumonia, and Normal categories. The study used a publicly available Mendeley Data dataset comprising 5,228 CXR images, split into 70% training, 10% validation, and 20% test sets. Three labeled data scenarios (5%, 10%, and 20%) were evaluated. AdaMatch leverages both labeled and unlabeled data through adaptive thresholding, distribution alignment, and consistency regularization. The best performance was achieved with 20% labeled data, achieving 98.19% accuracy, 98.23% sensitivity, and 98.24% F1-score. This performance closely approached that of the supervised baseline model, which achieved an accuracy of 98.95%, sensitivity of 98.98%, and an F1-score of 98.98%. These findings demonstrate that AdaMatch is an effective semi-supervised learning approach for improving CXR image classification under limited labeled data conditions.

Keywords: AdaMatch, Chest X-Ray, DenseNet-169, Multi-Class Classification, Semi-Supervised Deep Learning

Abstrak

Penyakit paru-paru, seperti Coronavirus Disease 2019 (COVID-19) dan pneumonia, masih menjadi tantangan kesehatan yang memerlukan diagnosis cepat dan akurat. Citra chest X-ray (CXR) banyak digunakan untuk mendukung diagnosis, namun interpretasinya masih bergantung pada radiolog dan ketersediaan data berlabel. Keterbatasan data berlabel menjadi kendala dalam pengembangan model deep learning berbasis supervised learning. Penelitian ini menerapkan pendekatan *semi-supervised deep learning* menggunakan AdaMatch dengan DenseNet-169 untuk klasifikasi multikelas citra CXR menjadi COVID-19, Pneumonia, dan Normal. Dataset publik Mendeley Data yang digunakan terdiri atas 5.228 citra CXR, dengan pembagian 70% data pelatihan, 10% validasi, dan 20% pengujian. Tiga skenario proporsi data berlabel, yaitu 5%, 10%, dan 20%, digunakan untuk mengevaluasi performa model. AdaMatch memanfaatkan data berlabel dan tidak berlabel melalui mekanisme adaptive thresholding, distribution alignment, dan consistency regularization. Hasil terbaik diperoleh pada skenario 20% data berlabel dengan akurasi 98,19%, sensitivitas 98,23%, dan F1-score 98,24%. Performa tersebut mendekati model supervised learning pembandingan yang memperoleh akurasi 98,95%, sensitivitas 98,98%, dan F1-score 98,98%. Temuan ini menunjukkan bahwa AdaMatch merupakan pendekatan *semi-supervised* yang efektif untuk meningkatkan klasifikasi citra CXR pada kondisi keterbatasan data berlabel.

Kata Kunci: AdaMatch, Chest X-Ray, DenseNet-169, Klasifikasi Multi-Kelas, Semi-Supervised Deep Learning

1. PENDAHULUAN

Penyakit paru-paru masih menjadi salah satu tantangan besar dalam sistem kesehatan global karena berkontribusi terhadap tingginya angka morbiditas dan mortalitas di berbagai negara [1]. Kasus penyakit seperti pneumonia dan COVID-19 menunjukkan bahwa gangguan pernapasan dapat berkembang dengan cepat menjadi kondisi serius apabila tidak dideteksi sejak dini [2]. Dalam praktik klinis, *Chest X-Ray* (CXR) merupakan salah satu modalitas pencitraan yang paling sering digunakan untuk membantu pemeriksaan paru karena proses pemeriksaannya relatif cepat, biaya lebih terjangkau, serta ketersediaan alat yang lebih luas dibandingkan *CT scan* [3]. Meskipun demikian, interpretasi citra CXR tidak selalu mudah karena beberapa kelainan paru memiliki pola abnormalitas yang samar, kontras rendah, dan kemiripan visual antar kelas penyakit [4]. Kondisi tersebut menyebabkan proses diagnosis masih sangat bergantung pada pengalaman radiolog dan berpotensi menimbulkan perbedaan interpretasi, terutama ketika jumlah pemeriksaan meningkat [5]. Oleh karena itu, pengembangan sistem klasifikasi otomatis berbasis *deep learning* mulai banyak dilakukan untuk membantu proses analisis citra medis secara lebih konsisten dan efisien.

Meskipun perkembangan *deep learning* telah menunjukkan hasil yang menjanjikan dalam klasifikasi citra medis, penerapannya pada citra CXR masih menghadapi tantangan yang cukup kompleks. Citra normal, COVID-19, dan pneumonia memiliki karakteristik radiologis yang saling berdekatan sehingga model membutuhkan representasi fitur yang kuat untuk membedakan pola antar kelas secara konsisten [4]. Tantangan tersebut semakin besar karena pengembangan model *deep learning* pada citra medis sangat bergantung pada data berlabel, sementara proses pelabelan citra memerlukan validasi tenaga ahli seperti radiolog yang membutuhkan waktu, biaya, dan sumber daya yang tidak sedikit [6]. Dalam praktiknya, banyak institusi kesehatan memiliki data citra dalam jumlah besar, tetapi hanya sebagian kecil yang telah memiliki label valid dan siap digunakan untuk pelatihan model [7]. Kondisi ini tidak hanya membatasi jumlah data pelatihan, tetapi juga berpotensi menyebabkan model mengalami *overfitting*, bias terhadap kelas tertentu, serta menurunkan kemampuan generalisasi pada data baru, terutama pada skenario dengan distribusi data yang tidak seimbang [8]. Dengan demikian, pendekatan *supervised learning* murni belum selalu memadai, terutama ketika jumlah data berlabel terbatas dan distribusi antarkelas tidak seimbang.

Penelitian terdahulu menunjukkan bahwa *Convolutional Neural Network* (CNN) telah menjadi fondasi utama dalam klasifikasi citra medis, terbukti efektif mendeteksi pneumonia dari citra CXR dengan akurasi mencapai 90.22% [4]. Salah satu arsitektur yang banyak digunakan adalah DenseNet, yang memiliki *dense connectivity* antar-layer sehingga mampu meningkatkan ekstraksi fitur dan mendukung *transfer learning*. Arsitektur DenseNet-169 konsisten menjadi arsitektur terbaik dibandingkan 21 arsitektur CNN lainnya, bahkan mencapai akurasi 98.15% dan *F1-score* 98.12% pada dataset COVIDx8B [9]. Secara khusus, DenseNet-169 terbukti unggul dalam klasifikasi COVID-19 maupun pneumonia, dengan akurasi mencapai 96.37% serta *recall* 98.89%, mengungguli AlexNet, ResNet-50, dan VGG [10]. Meskipun menunjukkan performa yang tinggi, capaian tersebut umumnya diperoleh pada kondisi data berlabel yang memadai. Pada implementasi dengan data klinis yang terbatas, performa model dilaporkan mengalami penurunan, dengan *recall* hanya mencapai 78,8% [11], mengindikasikan bahwa kemampuan generalisasi model masih sangat dipengaruhi oleh ketersediaan data berlabel. Kondisi ini menunjukkan bahwa meskipun *backbone* yang kuat berkontribusi signifikan terhadap performa klasifikasi, model *supervised* murni cenderung mengalami penurunan performa pada skenario data berlabel yang terbatas, sehingga diperlukan pendekatan pembelajaran yang tidak hanya bergantung pada arsitektur, tetapi juga mampu memanfaatkan data tidak berlabel secara lebih efektif [12].

Keterbatasan performa model *supervised* pada kondisi data berlabel terbatas mendorong berkembangnya pendekatan *semi-supervised learning* (SSL) yang memanfaatkan data tidak berlabel melalui mekanisme *pseudo-label*, yaitu pemberian label semu berdasarkan prediksi dengan probabilitas tertinggi untuk digunakan dalam proses pelatihan [3]. Berbagai penelitian menunjukkan efektivitas pendekatan ini, seperti MixMatch yang mampu mencapai akurasi 93% pada klasifikasi COVID-19 dari CXR dengan proporsi data berlabel yang kecil [13], serta FixMatch yang menunjukkan performa kompetitif dengan akurasi 72%, 74%, dan 76% masing-masing pada skenario 5%, 10%, dan 20% data berlabel [14]. Namun, sebagian besar metode SSL masih menggunakan *fixed confidence threshold* dalam seleksi *pseudo-label*, sehingga hanya prediksi dengan tingkat keyakinan tinggi yang dipertahankan sementara sampel lain yang berpotensi informatif diabaikan [14]. Keterbatasan ini menjadi lebih signifikan pada citra CXR yang memiliki kemiripan pola visual antar kelas, di mana meskipun deteksi kelas COVID-19 mencapai akurasi 98,1%, akurasi keseluruhan dapat menurun hingga 75,23% akibat kesulitan mendeteksi kelas pneumonia dan normal yang memiliki kemiripan pola visual [15]. Untuk mengatasi permasalahan tersebut, diperkenalkan konsep *distribution alignment* yang menyesuaikan distribusi prediksi data tidak berlabel agar lebih mendekati distribusi data berlabel, sehingga menghasilkan *pseudo-label* yang lebih representatif dan seimbang antar kelas [16].

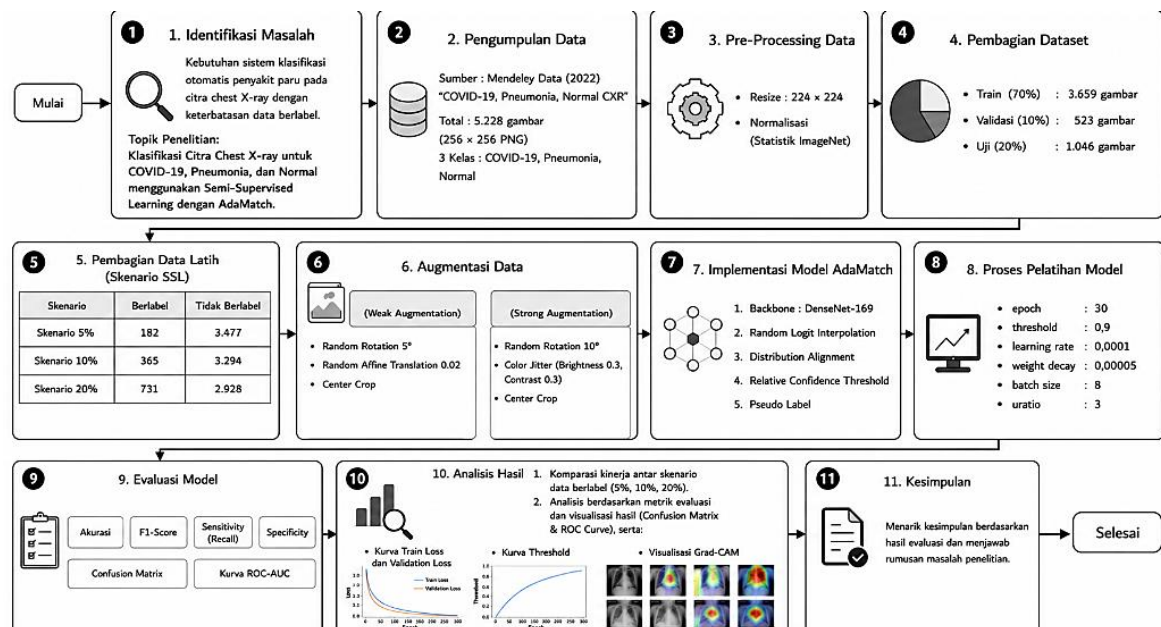
Salah satu metode yang dikembangkan untuk mengatasi permasalahan di atas adalah AdaMatch. Metode ini menggunakan mekanisme *adaptive confidence threshold* dan *distribution alignment* untuk meningkatkan kualitas *pseudo-label* selama proses pelatihan. Berbeda dengan pendekatan *pseudo-label* konvensional yang menggunakan *fixed confidence threshold*, AdaMatch menyesuaikan kriteria seleksi *pseudo-label* secara dinamis melalui *adaptive confidence threshold*, sementara *distribution alignment* digunakan untuk

menyeimbangkan distribusi probabilitas prediksi data tak berlabel agar konsisten dengan distribusi kelas pada data berlabel [17]. Penelitian tersebut melaporkan bahwa pada *semi-supervised learning*, AdaMatch mampu mencapai rata-rata akurasi 48,4%, lebih tinggi dibandingkan FixMatch yang hanya mencapai 46,3% [17]. Dalam konteks klasifikasi citra CXR, keunggulan tersebut menjadi penting karena karakteristik visual antar kelas sering memiliki kemiripan yang tinggi dan memerlukan kemampuan generalisasi model yang baik pada kondisi data berlabel terbatas. Integrasi AdaMatch dengan *backbone* DenseNet-169 juga berpotensi meningkatkan kemampuan model dalam mengekstraksi sekaligus mempelajari representasi fitur dari data berlabel maupun tidak berlabel secara lebih efektif.

Berdasarkan permasalahan yang telah diuraikan, penelitian ini bertujuan untuk menerapkan *semi-supervised deep learning* menggunakan AdaMatch dengan arsitektur dasar Densely Connected Convolutional Network-169 (DenseNet-169) dalam klasifikasi penyakit paru-paru pada citra *chest X-ray* (CXR). Penelitian ini dilakukan dengan menggunakan beberapa skenario proporsi data berlabel untuk mengevaluasi kemampuan model dalam kondisi keterbatasan label. Selain menganalisis performa klasifikasi pada kelas Normal, Pneumonia, dan Coronavirus Disease 2019 (COVID-19), penelitian ini juga mengevaluasi kemampuan AdaMatch dalam memanfaatkan data tidak berlabel selama proses pelatihan. Evaluasi model dilakukan menggunakan metrik akurasi, presisi, sensitivitas, spesifisitas, F1-score, confusion matrix, Receiver Operating Characteristic-Area Under the Curve (ROC-AUC), dan *Gradient-weighted Class Activation Mapping* (Grad-CAM) untuk memberikan analisis performa yang lebih menyeluruh. Melalui penelitian ini, diharapkan dapat diperoleh pemahaman mengenai efektivitas AdaMatch dalam mendukung pengembangan sistem klasifikasi citra medis yang adaptif terhadap keterbatasan data berlabel.

2. METODOLOGI PENELITIAN

Penelitian ini menggunakan pendekatan eksperimen berbasis *deep learning* untuk menerapkan dan mengevaluasi metode AdaMatch dengan *backbone* DenseNet-169 pada klasifikasi penyakit paru-paru menggunakan citra *chest x-ray* (CXR). Fokus penelitian diarahkan pada pengujian performa model pada kondisi keterbatasan data berlabel, sehingga proses eksperimen tidak hanya menilai hasil klasifikasi, tetapi juga kemampuan model dalam memanfaatkan data tidak berlabel selama pelatihan. Untuk mendukung pengujian tersebut, eksperimen dilakukan menggunakan tiga skenario proporsi data berlabel, yaitu 5%, 10%, dan 20% dari total data latih. Seluruh skenario diuji menggunakan dataset, *preprocessing*, *backbone*, parameter pelatihan, *train set*, *validation set*, dan *test set* yang sama agar perbandingan performa model dapat dilakukan secara lebih objektif. Alur tahapan penelitian secara keseluruhan dapat dilihat pada Gambar 1 yang menggambarkan proses penelitian secara sistematis. Visualisasi diagram didukung dengan bantuan alat berbasis AI generatif, namun tetap disesuaikan dengan alur dan metodologi penelitian yang dirancang.



Gambar 1. Metodologi Penelitian

Alur penelitian pada Gambar 1 dimulai dari identifikasi masalah dan pengumpulan dataset CXR, kemudian dilanjutkan dengan tahap *pre-processing* berupa *resize* citra dan normalisasi data. Dataset selanjutnya dibagi menjadi data latih, validasi, dan data uji menggunakan metode *stratified split* untuk menjaga distribusi kelas pada setiap subset data [18]. Setelah proses pembagian dataset, data latih dibentuk kembali ke

dalam beberapa skenario *semi-supervised learning* dengan proporsi data berlabel sebesar 5%, 10%, dan 20%, sedangkan sisa data digunakan sebagai data tidak berlabel selama proses pelatihan model. Tahap berikutnya meliputi augmentasi citra menggunakan *weak augmentation* dan *strong augmentation*, implementasi model AdaMatch dengan DenseNet-169, proses pelatihan model, evaluasi performa menggunakan beberapa metrik evaluasi, serta analisis hasil berdasarkan perbandingan antarskenario data berlabel.

2.1 Pengumpulan Dataset

Dataset yang digunakan pada penelitian ini merupakan dataset publik *Covid19-Pneumonia-Normal Chest X-ray Images* yang tersedia pada *platform Mendeley Data* [19]. Dataset dapat diakses melalui Covid19-Pneumonia-Normal Chest X-Ray Images - Mendeley Data. Dataset dipilih karena sesuai dengan tujuan penelitian, yaitu klasifikasi penyakit paru menggunakan citra CXR ke dalam tiga kelas, yaitu normal, COVID-19, dan pneumonia. Penggunaan dataset publik mendukung transparansi penelitian serta memudahkan peneliti lain untuk mereproduksi eksperimen. Selain itu, dataset ini juga telah digunakan pada beberapa penelitian sebelumnya, seperti CheXImageNet oleh Shastri dan LiteCovidNet oleh Kumar, sehingga memiliki relevansi akademik dalam pengembangan klasifikasi citra CXR berbasis *deep learning*[19] [20].



Gambar 2. Chest X-Ray Paru-Paru Normal, Pneumonia, dan COVID-19

Dataset tersusun dalam struktur folder berdasarkan kelas, sehingga setiap citra telah dikelompokkan sesuai kategori penyakitnya. Secara keseluruhan, dataset terdiri atas 5.228 citra CXR berformat PNG dengan ukuran 256×256 piksel. Distribusi data mencakup 1.802 citra Normal, 1.626 citra COVID-19, dan 1.800 citra Pneumonia. Komposisi tersebut menunjukkan bahwa dataset memiliki distribusi kelas yang relatif seimbang, sehingga cukup representatif untuk digunakan pada eksperimen klasifikasi multikelas serta mendukung penerapan *semi-supervised learning* pada kondisi data berlabel terbatas [14]. Meskipun demikian, perlu dicatat bahwa seluruh citra berasal dari satu sumber dataset publik, sehingga variasi klinis yang ditemukan di berbagai institusi kesehatan, perangkat akuisisi, maupun populasi pasien yang berbeda belum terwakili dalam penelitian ini. Keterbatasan ini berpotensi memengaruhi kemampuan generalisasi model apabila diterapkan pada data dari distribusi yang berbeda.

2.2 Preprocessing Data

Tahap *preprocessing* dilakukan untuk mempersiapkan citra sebelum digunakan dalam proses pelatihan model. Pada dataset CXR, citra dapat memiliki variasi ukuran dan distribusi nilai piksel yang berbeda, sehingga perlu dilakukan penyesuaian agar data memiliki karakteristik input yang lebih konsisten. *Preprocessing* juga membantu menjaga stabilitas proses pembelajaran model, terutama ketika menggunakan *backbone* DenseNet-169 dengan bobot *pretrained* [10]. Dengan input yang lebih terstandarisasi, model dapat mempelajari pola visual secara lebih optimal tanpa dipengaruhi oleh perbedaan teknis antarcitra.

Tahap *preprocessing* pada penelitian ini terdiri atas dua proses utama, yaitu *resize* citra dan *normalisasi*. Proses *resize* dilakukan dengan mengubah ukuran seluruh citra menjadi 224×224 piksel agar sesuai dengan ukuran input yang digunakan oleh DenseNet-169 [9]. Setelah proses *resize*, citra kemudian dinormalisasi untuk menyesuaikan distribusi nilai piksel. Normalisasi dilakukan menggunakan nilai *mean* dan *standard deviation* dari ImageNet, yaitu *mean* = [0,485; 0,456; 0,406] dan *standard deviation* = [0,229; 0,224; 0,225] [10]. Langkah ini diterapkan karena DenseNet-169 menggunakan bobot *pretrained* ImageNet, sehingga distribusi input perlu diselaraskan dengan karakteristik data pada tahap *pretraining*. Dengan *preprocessing* yang sederhana namun konsisten, model dapat menerima input yang lebih stabil selama proses pembelajaran [12].

2.3 Pembagian Data

Setelah melalui tahap *preprocessing*, dataset dibagi menjadi tiga subset utama, yaitu data latih (*train set*), data validasi (*validation set*), dan data uji (*test set*). Pembagian data dilakukan untuk memastikan proses pelatihan dan evaluasi model berlangsung secara terpisah, sehingga performa model dapat diukur secara lebih objektif [21]. Pada penelitian klasifikasi citra medis, pemisahan dataset menjadi beberapa subset penting untuk mengurangi risiko bias evaluasi dan menjaga agar model tidak diuji menggunakan data yang pernah dipelajari selama pelatihan [13].

Pembagian dataset dilakukan dengan rasio 70% untuk data latih, 10% untuk data validasi, dan 20% untuk data uji [21]. Proses pembagian dilakukan dengan *stratified split* agar distribusi kelas tetap terjaga pada setiap subset data [18]. Berdasarkan rasio tersebut, diperoleh 3.659 citra pada *train set*, 523 citra pada *validation set*, dan 1.046 citra pada *test set*. Data latih digunakan sebagai sumber utama pada proses pembelajaran model, sedangkan data validasi digunakan untuk memantau performa selama pelatihan dan membantu menentukan model terbaik (*robust*). Sementara itu, data uji digunakan pada tahap akhir penelitian untuk mengevaluasi kemampuan model terhadap data yang belum pernah dilihat sebelumnya. Pembagian ini diterapkan secara konsisten agar seluruh tahapan eksperimen dilakukan pada distribusi data yang sama [22].

2.4 Pembagian Data Berlabel dan Tidak Berlabel

Setelah pembagian dataset utama dilakukan, proses berikutnya difokuskan pada pembentukan skenario *semi-supervised learning* pada data latih. Pada tahap ini, data latih tidak langsung digunakan seluruhnya sebagai data berlabel, melainkan dipisahkan kembali menjadi subset data berlabel dan data tidak berlabel. Pemisahan ini dilakukan karena pendekatan *semi-supervised learning* memerlukan kombinasi kedua jenis data selama proses pelatihan [5][14]. Pada penelitian ini, kondisi data tidak berlabel dibentuk secara terkontrol pada subset data latih setelah proses pembagian dataset utama dilakukan [13]. Pendekatan ini digunakan untuk merepresentasikan situasi yang umum dijumpai pada pengembangan model medis, ketika jumlah citra tersedia cukup banyak, namun data dengan label valid masih terbatas. Pembentukan subset berlabel dan tidak berlabel dilakukan dengan beberapa skenario proporsi data berlabel, yaitu 5%, 10%, dan 20% dari total data latih [14].

Tabel 1. Skenario Data Berlabel dan Tidak Berlabel

Skenario	Data Berlabel	Data Tidak Berlabel	Total
Skenario 5%	182	3477	3659
Skenario 10%	365	3294	3659
Skenario 20%	731	2928	3659

Berdasarkan Tabel 1, pembagian data berlabel tetap mempertahankan informasi kelas asli, sedangkan sisa data pada *train set* diperlakukan sebagai data tidak berlabel tanpa menghilangkan informasi citra yang dimiliki. Strategi ini memungkinkan seluruh data tetap dimanfaatkan selama proses pelatihan, meskipun hanya sebagian kecil yang digunakan sebagai acuan label. Dengan demikian, dataset utama tetap berfungsi sebagai sumber data eksperimen, sementara pengaturan data berlabel terbatas digunakan untuk melihat bagaimana AdaMatch diterapkan dalam memanfaatkan data tidak berlabel pada klasifikasi penyakit paru [17].

2.5 Augmentasi Data

Augmentasi citra dilakukan pada tahap pelatihan untuk meningkatkan variasi data dan membantu model mempelajari representasi fitur yang lebih *robust*. Pada metode AdaMatch, proses augmentasi diterapkan pada data berlabel maupun data tidak berlabel menggunakan dua jenis transformasi, yaitu *weak augmentation* dan *strong augmentation* [17]. Pada penelitian ini, *weak augmentation* dirancang untuk mempertahankan karakteristik utama citra agar *pseudo-label* tetap stabil. Transformasi yang digunakan berupa *random rotation* sebesar 5°, *random affine translation* sebesar 0,02, *center crop*, konversi ke tensor, serta normalisasi menggunakan statistik ImageNet. Pemilihan kombinasi *weak augmentation* tersebut dilakukan untuk mempertahankan karakteristik anatomi paru pada citra CXR [5].

Sebaliknya, *strong augmentation* diterapkan untuk menghasilkan variasi visual yang lebih kompleks sehingga model dilatih agar konsisten terhadap perubahan distribusi data. Transformasi yang digunakan meliputi *random rotation* sebesar 10°, *color jitter* pada *brightness* dan *contrast* sebesar 0,3, *center crop*, konversi ke tensor, dan normalisasi berbasis ImageNet. Penggunaan rotasi 10° dipilih karena termasuk augmentasi geometris yang mampu meningkatkan keragaman sudut pandang citra, serta menunjukkan bahwa rotasi moderat lebih stabil dibandingkan rotasi ekstrem (misalnya 15° atau lebih) yang justru tidak memberikan peningkatan yang signifikan [23]. Selain itu, kombinasi *color jitter (brightness/contrast)* dan *cropping* menghasilkan representasi fitur yang lebih robust dan memiliki kemampuan generalisasi yang baik pada citra CXR [24]. Dengan demikian, kombinasi *weak* dan *strong augmentation* tidak hanya memperluas variasi data, tetapi juga menjaga keseimbangan antara stabilitas *pseudo-label* dan kemampuan generalisasi model [5].

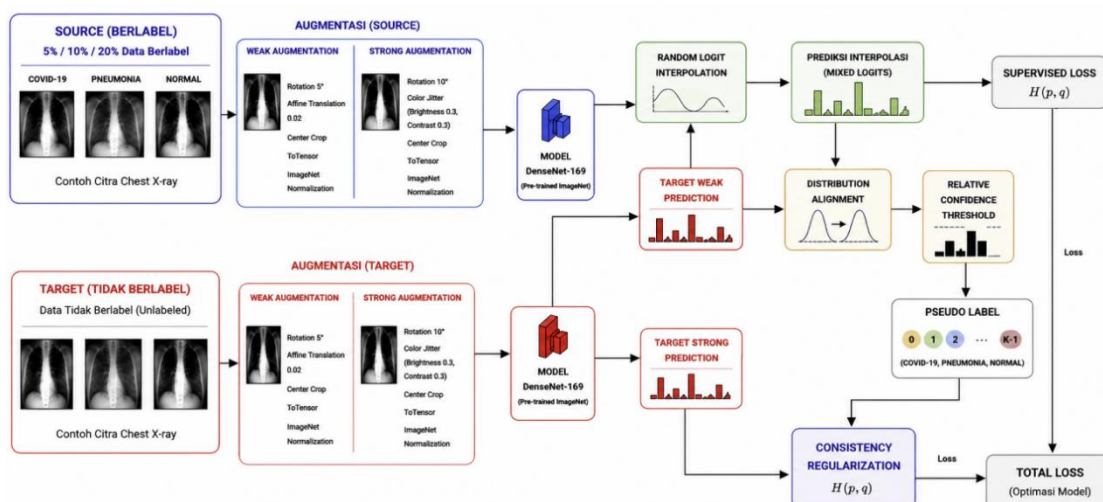
2.6 Penerapan Semi-Supervised Learning AdaMatch

Penelitian ini menerapkan metode *semi-supervised deep learning* menggunakan AdaMatch dengan DenseNet-169 sebagai *backbone* utama untuk klasifikasi citra CXR. Pemilihan DenseNet-169 didasarkan pada kemampuannya dalam mempertahankan aliran informasi antar *layer* melalui mekanisme *dense connectivity*, sehingga model mampu menghasilkan representasi fitur yang lebih kaya dan stabil pada citra medis [8]. Selain itu, penggunaan bobot *pretrained* ImageNet membantu mempercepat proses konvergensi dan meningkatkan kualitas ekstraksi fitur, terutama pada kondisi jumlah data berlabel yang terbatas [25]. Implementasi model pada penelitian ini mengadaptasi prinsip utama AdaMatch, yaitu *random logit interpolation*, *distribution*

alignment, *relative confidence thresholding*, dan *consistency regularization* untuk memanfaatkan data tidak berlabel secara lebih optimal [17].

Pada tahap implementasi, DenseNet-169 diinisialisasi menggunakan bobot *pretrained* dari ImageNet, kemudian lapisan *classifier* bawaan diganti dengan rangkaian *fully connected layer* yang disesuaikan dengan kebutuhan penelitian [25]. Struktur *classifier* diawali dengan menerima 1.664 fitur hasil ekstraksi *backbone* DenseNet-169. Fitur tersebut kemudian diproses oleh *layer* kedua yang berperan sebagai *hidden layer* dengan 512 neuron, berfungsi mengompres representasi 1.664 dimensi menjadi representasi yang lebih ringkas sekaligus mempelajari kombinasi fitur yang paling relevan untuk membedakan ketiga kelas penyakit paru. Setelah itu, *Batch Normalization* diterapkan untuk menstabilkan distribusi aktivasi antar *layer* sehingga proses pelatihan menjadi lebih stabil, terutama karena distribusi fitur dari *backbone pretrained* ImageNet belum tentu sesuai langsung dengan karakteristik domain citra CXR. Selanjutnya, Fungsi aktivasi ReLU digunakan agar model dapat mempelajari pola yang lebih kompleks dan tidak hanya bergantung pada hubungan linear antar fitur. *Dropout* dengan nilai 0,4 kemudian diterapkan sebagai mekanisme regularisasi dengan cara mematikan 40% neuron secara acak selama pelatihan, yang bertujuan mencegah model mengalami *overfitting*, terutama pada kondisi data berlabel yang sangat terbatas seperti pada skenario *semi-supervised learning*.

Terakhir, *layer* terakhir dengan 3 *neuron output* menghasilkan *logit* untuk masing-masing kelas, yaitu COVID-19, Pneumonia, dan Normal. Seluruh lapisan DenseNet-169, baik *backbone* maupun *classifier*, dilatih kembali melalui strategi *fine-tuning* dengan *learning rate* kecil sebesar $1e-4$. Pendekatan ini dipilih agar representasi fitur yang awalnya bersifat umum dari pelatihan ImageNet dapat menyesuaikan diri secara bertahap dengan pola radiologis khas citra CXR [25]. Setelah arsitektur model DenseNet-169 dikonfigurasi, proses pelatihan dilakukan menggunakan metode AdaMatch yang bekerja dengan memanfaatkan data berlabel dan data tidak berlabel secara bersamaan dalam proses pelatihan model. Secara umum, proses AdaMatch terdiri dari beberapa tahapan dapat dilihat pada Gambar 3. Alur kerja metode AdaMatch divisualisasikan dengan bantuan alat berbasis AI generatif dan disesuaikan dengan metode yang digunakan dalam penelitian ini.



Gambar 3. Alur Kerja Metode AdaMatch

1. Pengambilan Data Berlabel dan Tidak Berlabel

Pada setiap iterasi pelatihan, sistem mengambil dua jenis data, yaitu data berlabel (*source*) dan data tidak berlabel (*target*). Kedua data tersebut kemudian diproses masing-masing menggunakan dua jenis augmentasi, yaitu *weak augmentation* dan *strong augmentation*. Pada *data source*, hasil *weak augmentation* digunakan untuk menghitung rata-rata *confidence* dan menjadi bagian dari *Random Logit Interpolation*, sedangkan hasil *strong augmentation* langsung dibandingkan dengan label asli untuk menghitung *supervised loss*. Pada *data target*, hasil prediksi dari *weak augmentation* digunakan untuk membentuk *pseudo-label* dan kombinasi dari *random logit interpolation*, sedangkan hasil *strong augmentation* digunakan pada proses *consistency regularization* untuk perhitungan *unsupervised loss*.

2. Random Logit Interpolation

Random Logit Interpolation (RLI) dilakukan dengan mengombinasikan *logit* dari *source weak* dan *logit* gabungan *weak (source+target)* secara acak. Tujuannya adalah menstabilkan distribusi prediksi dan mengurangi perbedaan antara data berlabel dan tidak berlabel sebelum dilakukan penyesuaian distribusi, yang ditunjukkan pada persamaan 1.

$$P_{\text{new}} = \lambda \cdot P_{\text{source}} + (1 - \lambda) \cdot P_{\text{mix}} \quad (1)$$

3 Distribution alignment

Distribution alignment adalah proses penyesuaian distribusi prediksi data target agar mendekati distribusi data sumber. Tahapan ini penting untuk mengurangi bias model terhadap kelas tertentu selama pembentukan pseudo-label. Berikut persamaan distribution alignment yang digunakan pada AdaMatch, ditunjukkan pada persamaan 2.

$$\text{Normalize} = (q_u \times \frac{E[p1(y)]}{E[q_u]}) \quad (2)$$

4 Relative Confidence Thresholding

Relative confidence thresholding digunakan untuk menentukan apakah pseudo-label dari data tidak berlabel layak digunakan. Pseudo-label digunakan apabila nilai confidence lebih besar dari threshold adaptif yang telah ditentukan. Threshold adaptif ditentukan dengan menghitung confidence rata-rata dari data source, kemudian dikalikan dengan threshold default yang ditunjukkan pada persamaan 3.

$$\tau = \mu \times \frac{1}{B} \sum_{b=1}^B \max(p1(y_b)) \quad (3)$$

5 Consistency Regularization

Consistency regularization memastikan model menghasilkan prediksi yang konsisten antara target weak dan target strong. Pseudo-label dari target weak dijadikan acuan, lalu dibandingkan dengan prediksi target strong menggunakan cross-entropy untuk menghasilkan *loss unsupervised*. Persamaan unsupervised loss pada AdaMatch ditunjukkan pada persamaan 4

$$L_u = H(q_u, p_m(y | A_s(x_u))) \quad (4)$$

6 Perhitungan Loss

Pada AdaMatch, total loss merupakan kombinasi antara supervised loss (data source) dan unsupervised loss (data target). Persamaan total loss ditunjukkan pada persamaan 5.

$$L = L_s + \lambda_u L_u \quad (5)$$

di mana $\lambda_u(t)$ adalah fungsi warm-up yang meningkat seiring epoch. Kombinasi ini memungkinkan model belajar dari label asli sekaligus memanfaatkan informasi dari data tidak berlabel, sehingga lebih stabil, adaptif, dan robust dalam klasifikasi citra medis.

2.7 Pelatihan Model

Pada tahap proses pelatihan model, implementasi metode AdaMatch dilakukan menggunakan bahasa pemrograman Python dengan bantuan *library* PyTorch. *Optimizer* yang digunakan adalah Adam dengan *learning rate* 0.0001 dan *weight decay* 0.0005, yang dipilih untuk menjaga kestabilan proses pembelajaran serta mengurangi risiko *overfitting* pada data dengan jumlah terbatas. Penggunaan nilai *weight decay* yang relatif kecil juga sejalan dengan temuan yang menunjukkan bahwa regularisasi pada rentang rendah (10^{-4}) cenderung menghasilkan performa yang lebih stabil [17].

Pelatihan dilakukan selama 30 *epoch* dengan *batch size* 8, menyesuaikan dengan keterbatasan dataset medis dan memori GPU. Parameter *uratio* 3 digunakan untuk menyeimbangkan proporsi data berlabel dan tidak berlabel, di mana rasio ini dilaporkan mampu memberikan kontribusi yang signifikan terhadap peningkatan performa model dibandingkan dengan pemilihan rasio yang lebih rendah. Selain itu, digunakan *confidence threshold* 0.9, untuk memastikan bahwa *pseudo-label* yang digunakan memiliki tingkat kepercayaan yang cukup tinggi, namun tetap mempertahankan jumlah sampel yang memadai. Pemilihan nilai ini juga didasarkan pada rentang optimal (0,8-1.0) di mana performa cenderung menurun ketika nilai *threshold* terlalu mendekati 1,0 [17].

2.8 Evaluasi

Evaluasi model dilakukan menggunakan data uji yang tidak terlibat pada proses pelatihan maupun validasi, sehingga hasil pengujian dapat merepresentasikan kemampuan generalisasi model secara lebih objektif. Pada penelitian ini, performa AdaMatch dievaluasi menggunakan beberapa metrik, yaitu *accuracy*, *precision*, *sensitivity (recall)*, *F1-score*, *specificity*, *confusion matrix*, dan *ROC-AUC*. Penggunaan beberapa metrik evaluasi dilakukan karena klasifikasi citra medis tidak cukup dinilai dari satu indikator performa [26].

Accuracy digunakan untuk mengukur proporsi prediksi benar terhadap seluruh data uji, sedangkan *precision* digunakan untuk melihat ketepatan model dalam memberikan prediksi pada suatu kelas. *Sensitivity*

(recall) digunakan untuk mengukur kemampuan model dalam mendeteksi data yang benar-benar termasuk ke dalam kelas target, sehingga metrik ini penting pada klasifikasi penyakit untuk mengurangi kesalahan prediksi pasien sakit menjadi normal. Sementara itu, *F1-score* digunakan untuk melihat keseimbangan antara *precision* dan *sensitivity*, sedangkan *specificity* digunakan untuk mengukur kemampuan model dalam mengenali data negatif secara benar [26].

Analisis lebih rinci dilakukan menggunakan *confusion matrix* untuk melihat distribusi prediksi benar dan salah pada masing-masing kelas. Selain itu, penelitian ini juga menggunakan *Receiver Operating Characteristic (ROC)* dan *Area Under Curve (AUC)* untuk mengevaluasi kemampuan model dalam membedakan setiap kelas pada berbagai kelas penyakit [22]. Selain evaluasi performa akhir, penelitian ini turut menganalisis *train loss* dan *validation loss* untuk mengevaluasi stabilitas pembelajaran model. Perubahan *relative confidence threshold* (c_r) juga dianalisis selama proses pelatihan AdaMatch berlangsung untuk mengamati perkembangan tingkat kepercayaan model dalam menghasilkan *pseudo-label* secara adaptif pada setiap *epoch*. Analisis kualitas *pseudo-label* juga dilakukan dengan mengevaluasi persentase *pseudo-label* yang diterima, akurasi *pseudo-label* terhadap label asli, serta distribusi *pseudo-label* antar kelas, guna memastikan bahwa data tidak berlabel dimanfaatkan secara optimal dan tidak menimbulkan bias selama proses pelatihan. Sebagai pendukung interpretabilitas visual, penelitian ini juga menggunakan *Gradient-weighted Class Activation Mapping* (Grad-CAM) untuk menampilkan area citra CXR yang paling memengaruhi keputusan model saat proses klasifikasi [27].

3. HASIL DAN PEMBAHASAN

3.1. Hasil Evaluasi Kinerja Model

Penelitian ini menerapkan metode *semi-supervised deep learning* menggunakan AdaMatch dengan *backbone* DenseNet-169 untuk klasifikasi citra *chest x-ray (CXR)* ke dalam tiga kelas, yaitu COVID-19, pneumonia, dan normal. Pengujian dilakukan menggunakan tiga skenario jumlah data berlabel, yaitu 5%, 10%, dan 20%, sedangkan sisa data digunakan sebagai data tidak berlabel dalam proses *semi-supervised learning* (SSL). Untuk mengevaluasi kontribusi nyata dari pendekatan SSL, hasil AdaMatch dibandingkan secara langsung dengan model *supervised* DenseNet-169. Kedua model dilatih menggunakan *backbone*, parameter pelatihan, serta pembagian data berlabel yang sama pada setiap skenario, sehingga perbandingan performa dapat dilakukan secara objektif.

Tabel 2. Perbandingan Kinerja Model AdaMatch dan *Supervised* DenseNet-169

Skenario	Metode	Accuracy	Precision	Recall	Specificity	F1-score
100% Labeled	<i>Supervised</i> DenseNet-169	98,95 %	98,98 %	98,98 %	99,47 %	98,98 %
5% Labeled	SSL AdaMatch	96,08%	96,24%	96,20%	98,01%	96,19%
10% Labeled	SSL AdaMatch	96,85%	97,03%	96,95%	98,40%	96,93%
20% Labeled	SSL AdaMatch	98,19%	98,25%	98,23%	99,08%	98,24%

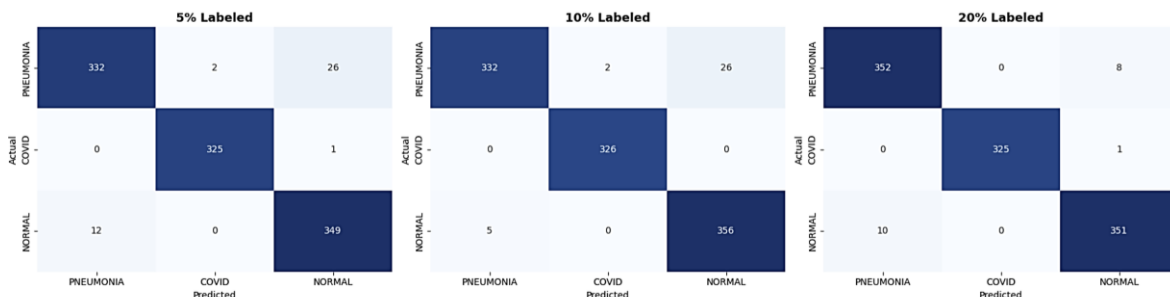
Berdasarkan Tabel 2, performa model AdaMatch menunjukkan peningkatan yang konsisten seiring bertambahnya proporsi data berlabel, dari *accuracy* 96,08% pada skenario 5% *labeled* menjadi 98,19% pada skenario 20% *labeled*. Peningkatan ini diikuti pula oleh metrik *recall* yang naik dari 96,20% menjadi 98,23%, yang dalam konteks klasifikasi citra medis menjadi indikator penting karena berkaitan langsung dengan kemampuan model dalam mengurangi kasus *false negative*, yaitu kondisi pasien yang sebenarnya sakit namun terklasifikasi sebagai normal. Konsistensi peningkatan pada seluruh metrik evaluasi ini menunjukkan bahwa penambahan data berlabel membantu model AdaMatch membentuk representasi fitur yang semakin stabil, sehingga mekanisme *distribution alignment* dan *relative confidence threshold* dapat bekerja lebih optimal dalam menghasilkan *pseudo-label* berkualitas pada data tidak berlabel.

Analisis juga dilakukan dengan membandingkan performa AdaMatch dengan model *supervised* DenseNet-169 yang dilatih menggunakan 100% data berlabel. Model *supervised* tersebut mencapai *accuracy* sebesar 98,95% dan *F1-score* sebesar 98,98%, yang secara teoritis merepresentasikan batas atas (*upper bound*) performa yang dapat dicapai pada dataset ini. Sementara itu, hasil AdaMatch pada skenario 20% *labeled*, yaitu *accuracy* 98,19% dan *F1-score* 98,24%, hanya berselisih 0,76% pada *accuracy* dan 0,74% pada *F1-score* dibandingkan *upper bound*. Selisih yang sangat kecil ini mengindikasikan bahwa AdaMatch mampu mendekati performa optimal model meskipun hanya menggunakan seperlima dari data berlabel. Temuan tersebut menegaskan efektivitas mekanisme *pseudo-labeling* dan *consistency regularization* dalam mengompensasi keterbatasan data berlabel pada domain *medical imaging*.

Secara keseluruhan, hasil ini memperkuat argumen bahwa kombinasi AdaMatch dan DenseNet-169 efektif dalam mengekstraksi pola radiologis pada citra CXR melalui mekanisme *dense connectivity* antar-layer, sekaligus memanfaatkan informasi dari data tidak berlabel secara optimal. Dengan demikian, pendekatan ini berpotensi besar untuk diterapkan pada kondisi nyata di lapangan, di mana ketersediaan data medis berlabel sering menjadi kendala utama dalam pengembangan model *deep learning* untuk diagnosis penyakit paru berbasis citra CXR.

3.2. Analisis Confusion Matrix

Analisis *confusion matrix* dilakukan untuk melihat distribusi prediksi model pada masing-masing kelas secara lebih rinci. Melalui *confusion matrix*, dapat diamati jumlah prediksi yang benar maupun salah pada kelas Pneumonia, COVID-19, dan Normal untuk setiap skenario data berlabel. Analisis ini penting karena dalam klasifikasi citra medis, kesalahan prediksi dapat memengaruhi interpretasi kondisi pasien, terutama ketika citra penyakit diprediksi sebagai normal (*false negative*).



Gambar 4. Confusion matrix dengan AdaMatch

Skenario 5% *labeled* pada Gambar 4 menunjukkan bahwa sebagian besar citra berhasil diklasifikasikan dengan benar oleh model. Kelas COVID-19 menunjukkan performa yang sangat baik, dengan 325 citra terklasifikasi dengan benar dan hanya 1 citra yang salah diprediksi sebagai normal. Sementara itu, kesalahan prediksi paling banyak terjadi pada kelas pneumonia, di mana 26 citra diprediksi sebagai normal dan 2 citra diprediksi sebagai COVID-19. Selain itu, terdapat 12 citra normal yang salah diprediksi sebagai pneumonia. Hasil tersebut menunjukkan bahwa pada kondisi data berlabel yang sangat terbatas, model masih mengalami kesulitan membedakan beberapa pola visual antara pneumonia dan kondisi normal. Kesulitan ini muncul karena ciri radiologis pneumonia sering kali menyerupai variasi paru normal, terutama pada citra dengan kualitas dan kontras rendah.

Pada skenario 10% *labeled*, distribusi prediksi terlihat lebih stabil. Kelas COVID-19 meningkat menjadi 326 citra benar tanpa kesalahan terhadap pneumonia, sementara kelas normal mencapai 356 citra benar dengan hanya 5 kesalahan terhadap pneumonia. Meski demikian, kelas pneumonia masih menunjukkan pola kesalahan yang serupa, yaitu 26 citra diprediksi sebagai Normal dan 2 citra sebagai COVID-19. Hal ini menegaskan bahwa meskipun penambahan data berlabel membantu memperbaiki representasi fitur, karakteristik visual pneumonia tetap menjadi kelas yang paling menantang karena overlap intensitas dan tekstur dengan kelas normal.

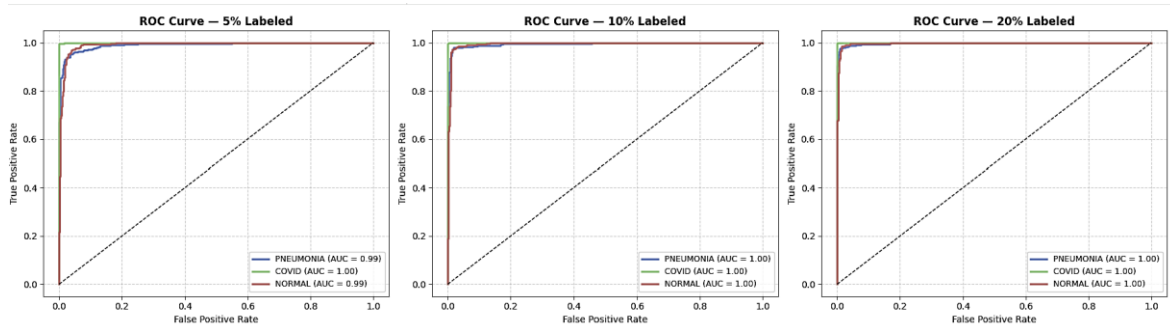
Performa terbaik diperoleh pada skenario 20% *labeled*, di mana kesalahan klasifikasi pneumonia menurun secara signifikan menjadi hanya 8 citra yang diprediksi sebagai normal. Kelas COVID-19 tetap konsisten dengan 325 prediksi benar dan hanya 1 kesalahan ke normal, sementara kelas normal mencapai 351 prediksi benar dengan 10 kesalahan ke pneumonia. Penurunan kesalahan ini menunjukkan bahwa penambahan data berlabel memperkaya representasi fitur dalam lapisan DenseNet-169, sehingga model lebih mampu membedakan pola radiologis Pneumonia dari variasi normal paru. Dengan kata lain, semakin banyak data berlabel yang tersedia, semakin tajam *feature separation* yang terbentuk di ruang laten, sehingga *decision boundary* antarkelas menjadi lebih jelas dan stabil.

3.3. Analisis ROC Curve

Analisis *Receiver Operating Characteristic* (ROC) dan *Area Under Curve* (AUC) dilakukan untuk mengevaluasi kemampuan model dalam membedakan setiap kelas pada berbagai penyakit. Kurva ROC menunjukkan hubungan antara *True Positive Rate* (TPR) dan *False Positive Rate* (FPR), sedangkan nilai AUC digunakan untuk mengukur kualitas kemampuan diskriminatif model. Semakin mendekati nilai 1, semakin baik kemampuan model dalam membedakan antarkelas. Kurva ROC dapat ditunjukkan pada Gambar 5.

Skenario 5% *labeled* pada Gambar 5 menunjukkan performa ROC yang sangat baik dengan nilai AUC sebesar 0,99 untuk kelas Pneumonia dan Normal, serta 1,00 untuk kelas COVID-19. Hasil tersebut menunjukkan bahwa meskipun jumlah data berlabel sangat terbatas, model masih mampu membedakan ketiga kelas secara efektif. Kurva ROC yang berada dekat dengan sudut kiri atas juga menunjukkan bahwa model memiliki sensitivitas tinggi dengan tingkat kesalahan prediksi yang rendah. Pada skenario 10% *labeled* dan 20% *labeled*, performa ROC-AUC mengalami peningkatan dibandingkan skenario sebelumnya. Seluruh kelas menunjukkan nilai AUC mendekati 1,00, yang menandakan bahwa model semakin baik dalam memisahkan distribusi antarkelas. Peningkatan ini menunjukkan bahwa penambahan jumlah data berlabel membantu model membangun representasi fitur yang lebih stabil selama proses *semi-supervised learning*. Secara umum, hasil

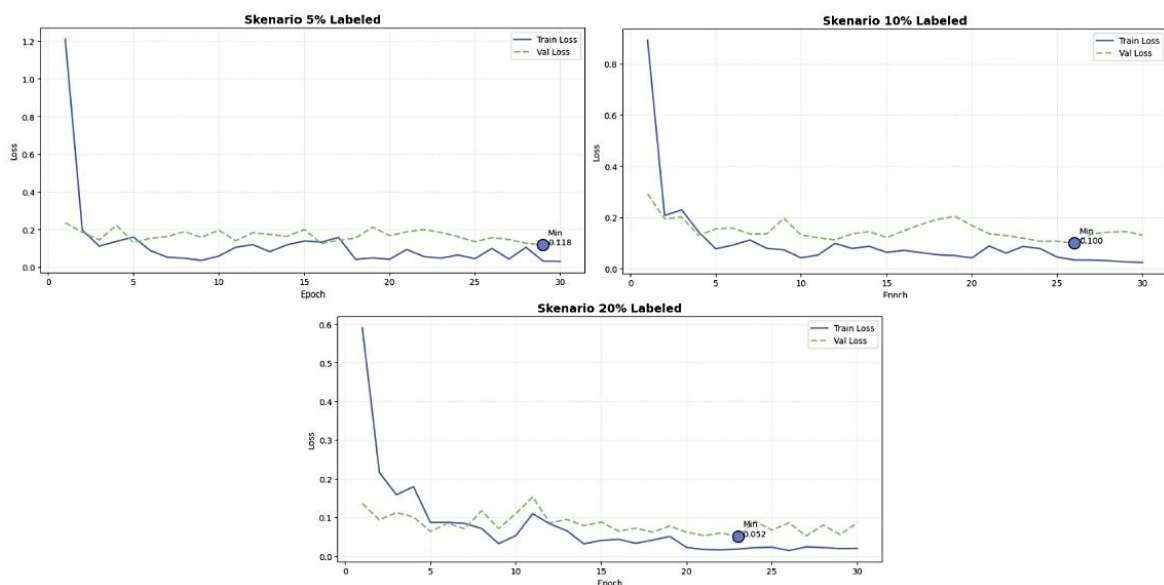
ROC-AUC pada seluruh skenario menunjukkan bahwa penerapan AdaMatch dengan DenseNet-169 mampu menghasilkan kemampuan klasifikasi yang sangat baik, bahkan pada kondisi data berlabel yang terbatas.



Gambar 5. Kurva ROC

3.4. Analisis Train loss dan Validation Loss

Analisis *train loss* dan *validation loss* dilakukan untuk mengevaluasi stabilitas proses pelatihan serta kemampuan generalisasi model pada setiap skenario proporsi data berlabel. Kurva *loss* memberikan gambaran bagaimana model mempelajari pola data selama proses pelatihan, sekaligus mengindikasikan potensi *overfitting* maupun *underfitting*. Pada penelitian ini, proses pelatihan dilakukan selama 30 *epoch* untuk setiap skenario, baik pada pendekatan *semi-supervised learning* (AdaMatch). Plot Train loss dan Validation Loss Model AdaMatch ditunjukkan pada Gambar 6.



Gambar 6. Plot Train loss dan Validation Loss Model AdaMatch

Skenario 5% *labeled* pada Gambar 6 menunjukkan nilai *train loss* mengalami penurunan yang cukup tajam pada awal pelatihan, kemudian cenderung stabil pada *epoch* berikutnya. Kondisi tersebut menunjukkan bahwa model mampu mempelajari pola dasar citra CXR dengan cepat meskipun jumlah data berlabel relatif sedikit. Sementara itu, *validation loss* terlihat berfluktuasi ringan selama proses pelatihan, namun tetap berada dalam rentang yang stabil tanpa peningkatan yang signifikan. Nilai *validation loss* terendah diperoleh pada *epoch* ke-29 sebesar 0,118, yang menunjukkan bahwa model masih mampu mempertahankan kemampuan generalisasi dengan baik pada kondisi label yang terbatas.

Pada skenario 10% *labeled*, pola penurunan *train loss* terlihat lebih stabil dibandingkan skenario sebelumnya. Nilai *validation loss* juga menunjukkan fluktuasi yang lebih kecil dengan kecenderungan menurun secara bertahap hingga akhir pelatihan. Kondisi ini menunjukkan bahwa penambahan jumlah data berlabel membantu model membentuk representasi fitur yang lebih konsisten selama proses *semi-supervised learning*. Nilai *validation loss* minimum pada skenario ini diperoleh pada *epoch* ke-26 sebesar 0,100, yang mengindikasikan peningkatan stabilitas model dibandingkan skenario 5% *labeled*.

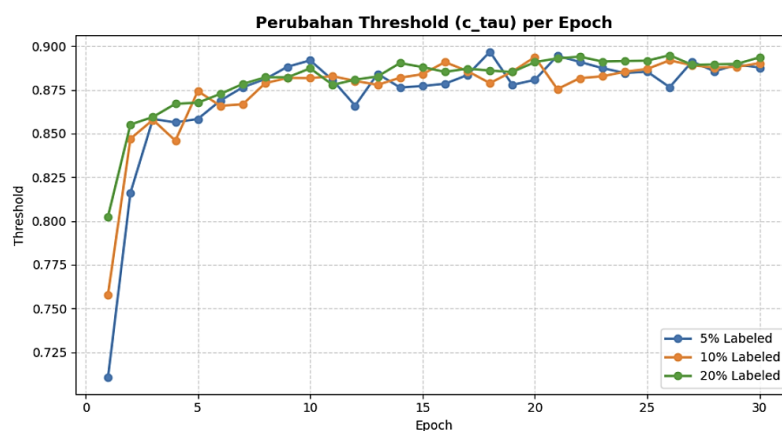
Performa pelatihan terbaik terlihat pada skenario 20% *labeled*. Pada skenario ini, *train loss* dan *validation loss* sama-sama menunjukkan pola penurunan yang lebih halus dan stabil selama proses pelatihan.

Jarak antara kedua kurva juga relatif kecil, sehingga menunjukkan bahwa model tidak mengalami *overfitting* yang signifikan. Nilai *validation loss* minimum diperoleh pada *epoch* ke-23 sebesar 0,052, yang merupakan nilai terendah dibandingkan seluruh skenario pengujian. Hasil tersebut menunjukkan bahwa penambahan data berlabel membantu model melakukan proses pembelajaran secara lebih optimal.

Secara umum, seluruh skenario menunjukkan pola pelatihan yang relatif stabil tanpa lonjakan *validation loss* yang ekstrem. Hal ini mengindikasikan bahwa kombinasi AdaMatch dan DenseNet-169 berpotensi menjaga konsistensi proses pembelajaran pada berbagai proporsi data yang berlabel. Selain itu, penggunaan *distribution alignment*, *relative confidence threshold*, dan kombinasi *weak* dan *strong augmentation* pada AdaMatch turut membantu model memanfaatkan data *unlabeled* secara lebih efektif selama proses pelatihan berlangsung.

3.5. Perubahan Relative Confidence Threshold

Analisis perubahan *relative confidence threshold* dilakukan untuk melihat dinamika *confidence* model selama proses pelatihan AdaMatch berlangsung. Pada penelitian ini, *threshold* tidak menggunakan nilai tetap (*fixed confidence threshold*), tetapi dihitung secara *adaptive* berdasarkan rata-rata *confidence prediction* pada data berlabel di setiap *batch*. Mekanisme tersebut memungkinkan proses pemilihan *pseudo-label* menyesuaikan dengan kondisi pembelajaran model pada setiap *epoch*.



Gambar 7. Kurva Relative Confidence Threshold

Berdasarkan grafik pada Gambar 7, seluruh skenario menunjukkan pola peningkatan *threshold* seiring bertambahnya *epoch* pelatihan. Pada awal *training*, nilai *threshold* cenderung lebih rendah karena *confidence* model terhadap prediksi masih belum stabil. Namun, setelah beberapa *epoch*, *threshold* mulai meningkat secara bertahap yang menunjukkan bahwa model mulai menghasilkan prediksi dengan tingkat keyakinan yang lebih tinggi. Pada skenario 5% *labeled*, perubahan *threshold* terlihat lebih fluktuatif dibandingkan skenario lainnya. Kondisi ini menunjukkan bahwa ketika jumlah data berlabel sangat terbatas, *confidence* model masih mengalami perubahan yang cukup dinamis selama proses pembelajaran. Meskipun demikian, *threshold* tetap menunjukkan tren meningkat hingga akhir pelatihan, yang menandakan bahwa model tetap mampu membangun *pseudo-label* yang semakin stabil seiring bertambahnya *epoch*.

Pada skenario 10% dan 20% *labeled*, pola perubahan *threshold* terlihat lebih stabil dengan fluktuasi yang lebih kecil. Hal ini menunjukkan bahwa penambahan jumlah data berlabel membantu model menghasilkan *confidence prediction* yang lebih konsisten selama proses pelatihan. Selain itu, *threshold* pada skenario 20% *labeled* cenderung mencapai nilai lebih tinggi dibandingkan dua skenario lainnya, yang mengindikasikan bahwa model menjadi lebih selektif dalam memilih *pseudo-label* untuk digunakan pada proses *unsupervised learning*. Pendekatan ini mampu menyesuaikan tingkat *confidence pseudo-label* secara dinamis selama proses pelatihan berlangsung. Dengan mekanisme tersebut, AdaMatch dapat memanfaatkan data *unlabeled* secara lebih fleksibel dibandingkan metode FixMatch yang menggunakan *fixed threshold*.

3.6. Analisis Dinamika Pseudo-Label

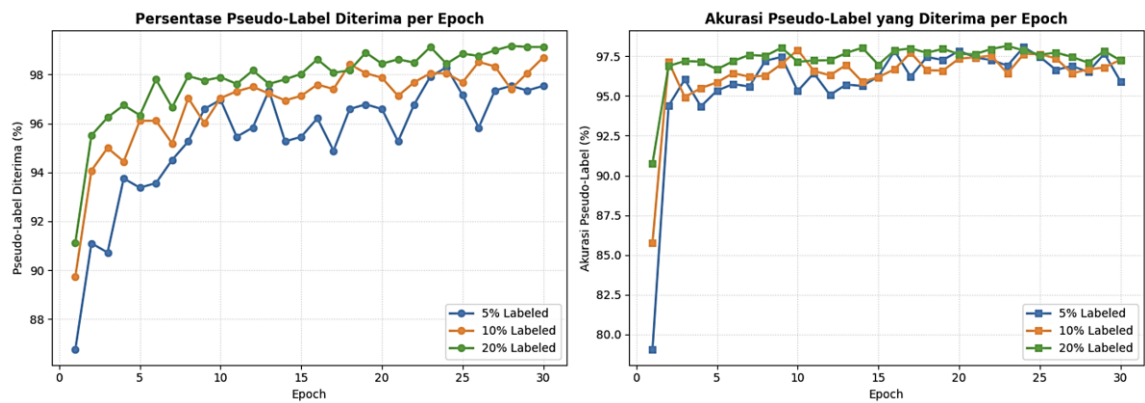
Analisis kualitas *pseudo-label* dilakukan untuk mengevaluasi sejauh mana mekanisme *adaptive threshold* dan *distribution alignment* pada AdaMatch mampu menghasilkan *pseudo-label* yang akurat dan representatif selama proses pelatihan. Evaluasi ini penting karena performa akhir model SSL sangat bergantung pada kualitas *pseudo-label* yang digunakan pada proses *consistency regularization*. Analisis dilakukan dengan menghitung tiga aspek, yaitu persentase *pseudo-label* yang diterima (lolos *relative confidence threshold*), akurasi *pseudo-label* terhadap label asli, serta distribusi *pseudo-label* pada setiap kelas.

Berdasarkan Tabel 3, persentase *pseudo-label* yang diterima meningkat seiring bertambahnya proporsi data berlabel, yaitu dari 97,54% pada skenario 5% *labeled* menjadi 99,13% pada skenario 20% *labeled*. Hasil

ini menunjukkan bahwa mekanisme *relative confidence thresholding* pada AdaMatch berhasil mempertahankan proporsi data tidak berlabel yang tinggi untuk dimanfaatkan dalam proses pelatihan. Meskipun tingkat penerimaan *pseudo-label* tinggi, akurasi *pseudo-label* tetap terjaga pada rentang 95,92%–97,28% di seluruh skenario, yang mengindikasikan bahwa *pseudo-label* yang dihasilkan tidak hanya banyak secara kuantitas, tetapi juga berkualitas. Selain itu, distribusi *pseudo-label* antar kelas relatif seimbang, dengan proporsi pada kisaran 30–36% untuk setiap kelas, sehingga model tidak menunjukkan bias terhadap kelas tertentu selama pembentukan *pseudo-label* berlangsung.

Tabel 3. Kualitas *Pseudo-label*

Skenario	<i>Pseudo-label</i> Diterima	Akurasi <i>Pseudo-label</i>	Distribusi Kelas		
			Pneumonia	COVID-19	Normal
5%	97,54 %	95,92 %	36,1 %	30,1 %	33,8 %
10%	98,70 %	97,28 %	34,1 %	30,7 %	35,3 %
20%	99,13 %	97,23 %	33,6 %	31,8 %	34,5 %

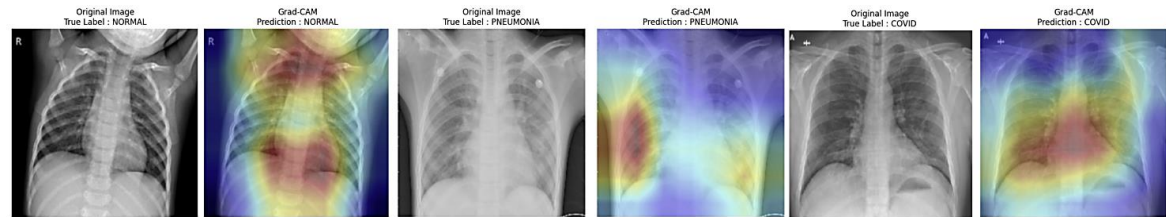


Gambar 8. Perubahan Persentase dan Akurasi *Pseudo-label* per Epoch

Gambar 8 menunjukkan dinamika *pseudo-label* selama proses pelatihan. Pada *epoch* awal pelatihan, persentase *pseudo-label* yang diterima masih relatif rendah, yaitu 86,77% hingga 91,12% pada seluruh skenario, dengan akurasi *pseudo-label* yang juga rendah, terutama pada skenario 5% *labeled* yang hanya mencapai 79,04% pada *epoch* pertama. Kondisi ini terjadi karena model belum memiliki representasi fitur yang memadai akibat proses pembelajaran yang terbatas pada tahap awal pelatihan. Namun, baik persentase penerimaan maupun akurasi *pseudo-label* meningkat secara signifikan setelah beberapa *epoch* awal dan cenderung stabil pada rentang 96%–99% hingga akhir pelatihan. Pola tersebut mengindikasikan bahwa mekanisme *adaptive confidence threshold* pada AdaMatch mampu menyesuaikan kriteria penerimaan *pseudo-label* secara dinamis seiring meningkatnya kepercayaan model terhadap prediksinya, sehingga *pseudo-label* yang dihasilkan tetap terjaga meskipun jumlah yang diterima terus bertambah.

3.7. Interpretabilitas Visual Menggunakan Grad-CAM

Analisis *Gradient-weighted Class Activation Mapping* (Grad-CAM) dilakukan untuk melihat area pada citra CXR yang paling memengaruhi keputusan model selama proses klasifikasi. Visualisasi Grad-CAM ditampilkan dalam bentuk *heatmap*, di mana warna merah dan kuning menunjukkan area dengan kontribusi paling tinggi terhadap prediksi model, sedangkan warna biru menunjukkan area dengan kontribusi yang lebih rendah. Analisis ini digunakan untuk membantu memahami apakah model benar-benar memfokuskan perhatian pada area paru-paru yang relevan secara visual, bukan pada area di luar objek utama. Selain itu, visualisasi Grad-CAM juga membantu memberikan interpretasi terhadap proses pembelajaran model yang umumnya bersifat *black box* dalam *deep learning*. Visualisasi Grad-CAM dapat ditunjukkan pada Gambar 9.



Gambar 9. Visualisasi Grad-CAM

Penggunaan Grad-CAM dalam penelitian ini mengacu pada pendekatan *visual explanation* berbasis gradien untuk menelusuri area yang berpengaruh terhadap keputusan model, yang diterapkan pada data uji (*test set*) untuk setiap kelas penyakit. Grad-CAM pada penelitian ini berfungsi sebagai pendukung interpretasi hasil klasifikasi, bukan sebagai bukti klinis secara langsung. Visualisasi yang dihasilkan menunjukkan bahwa perhatian model sebagian besar berada pada area anatomis yang wajar, meskipun makna klinis dari area aktivasi tersebut tetap memerlukan penilaian tenaga ahli radiologi. Oleh karena itu, hasil Grad-CAM perlu dibaca bersama metrik kuantitatif, *confusion matrix*, dan analisis per kelas agar interpretasi performa model tetap proporsional.

4. KESIMPULAN

Penelitian ini menerapkan AdaMatch dengan backbone DenseNet-169 untuk klasifikasi citra chest X-ray (CXR) ke dalam tiga kelas penyakit paru pada tiga skenario data berlabel sebesar 5%, 10%, dan 20%. Hasil eksperimen menunjukkan bahwa mekanisme *adaptive confidence threshold* pada AdaMatch dapat memanfaatkan data tidak berlabel secara efektif, dibuktikan melalui perbandingan langsung dengan model supervised DenseNet-169. AdaMatch secara konsisten mendekati performa model supervised pada seluruh skenario, dengan hasil terbaik pada skenario 20% labeled (accuracy 98,19% dan F1-score 98,24%). Performa tersebut hanya berselisih 0,76% pada accuracy dibandingkan dengan model supervised dengan 100% data berlabel, mengindikasikan bahwa AdaMatch mampu mendekati performa model *supervised baseline*, meskipun hanya menggunakan sebagian kecil data berlabel. Kesalahan klasifikasi paling banyak terjadi antara kelas Pneumonia dan Normal akibat kemiripan pola visual, sedangkan kelas COVID-19 cenderung lebih mudah dikenali oleh model. Hasil ROC-AUC mendekati 1,00 serta kurva loss yang stabil mendukung bahwa model memiliki kemampuan diskriminatif yang baik tanpa indikasi *overfitting* yang signifikan.

Komponen utama AdaMatch menunjukkan bahwa mekanisme *adaptive thresholding* mampu mempertahankan tingkat penerimaan *pseudo-label* yang tinggi (97,54%–99,13%) tanpa mengorbankan akurasi yang tetap terjaga pada rentang 95,92%–97,28%, dengan distribusi antarkelas yang relatif seimbang. Hal ini menegaskan fleksibilitas AdaMatch dibandingkan dengan pendekatan *fixed confidence threshold* pada metode SSL konvensional seperti FixMatch. Namun, evaluasi penelitian ini masih terbatas pada analisis deskriptif terhadap perilaku masing-masing komponen secara terpisah, sehingga kontribusi kuantitatif dari mekanisme seperti *distribution alignment*, *adaptive thresholding*, dan *consistency regularization* terhadap peningkatan performa akhir belum dapat diidentifikasi secara pasti. Analisis ablasi (*ablation study*) diperlukan untuk memastikan kontribusi nyata dari setiap komponen. Selain itu, visualisasi Grad-CAM menunjukkan fokus model pada area anatomis paru yang relevan, tetapi interpretasi ini masih bersifat komputasional dan belum divalidasi oleh ahli radiologi, sehingga keterlibatan radiolog menjadi penting untuk menilai relevansi diagnostik secara klinis.

Penelitian ini juga membuka peluang untuk pengembangan lebih lanjut. Penggunaan satu dataset membatasi variasi data yang digunakan, sehingga penelitian selanjutnya perlu menguji model pada dataset dari domain berbeda untuk memastikan konsistensi performa meskipun karakteristik data berubah. Selain itu, klasifikasi pada penelitian ini masih terbatas pada tiga kelas, penelitian berikutnya dapat menambahkan penyakit paru lain, seperti tuberkulosis atau kanker paru, untuk memperluas ruang lingkup klasifikasi. Lebih lanjut, pengujian menggunakan data pasien real-world dari rumah sakit serta integrasi model ke dalam Clinical Decision Support System (CDSS) diperlukan untuk menilai efektivitas model pada penerapan klinis di lapangan.

REFERENSI

- [1] WHO, "Pneumonia in children," 11 November 2022. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/pneumonia>
- [2] WHO, "WHO COVID-19 dashboard," July 2023. [Online]. Available: <https://data.who.int/dashboards/covid19/deaths>
- [3] M. Sakaida *et al.*, "The Effectiveness of Semi-Supervised Learning Techniques in Identifying Calcifications in X-ray Mammography and the Impact of Different Classification Probabilities," *Appl. Sci.*, vol. 14, no. 14, 2024, doi: 10.3390/app14145968.
- [4] A. Al Foysal and S. Sultana, "AI-Driven Pneumonia Diagnosis Using Deep Learning : A Comparative Analysis of CNN Models on Chest X-Ray Images," vol. 12, 2025, doi: 10.4236/oalib.1112899.
- [5] P. Sahoo, I. Roy, R. Ahlawat, S. Irtiza, and L. Khan, "Potential diagnosis of COVID-19 from chest X-ray and CT findings using semi-supervised learning," *Phys. Eng. Sci. Med.*, vol. 45, no. 1, pp. 31–42, 2022, doi: 10.1007/s13246-021-01075-2.
- [6] J. Yang, M. Chen, Q. Jia, and S. Liu, "Unlabeled Insight , Labeled Boost : Contrastive Learning and Class-Adaptive Pseudo-Labeling for Semi-Supervised Medical Image Classification," pp. 1–20, 2025, doi: doi.org/10.3390/e27101015.
- [7] J. Zhou, B. Jing, Z. Wang, H. Xin, and H. Tong, "SODA: Detecting COVID-19 in Chest X-Rays With Semi-Supervised Open Set Domain Adaptation," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol.

- 19, no. 5, pp. 2605–2612, 2022, doi: 10.1109/TCBB.2021.3066331.
- [8] G. Huang, Q. Fu, M. Gu, N. Lu, K. Liu, and T. Chen, “Deep Transfer Learning for the Multilabel Classification of Chest X-ray Images,” 2022, doi: doi.org/10.3390/diagnostics12061457.
- [9] F. A. Breve, “COVID-19 detection on Chest X-ray images : A comparison of CNN architectures and ensembles,” *Expert Syst. Appl.*, vol. 204, no. May, p. 117549, 2022, doi: 10.1016/j.eswa.2022.117549.
- [10] P. Pradeep, D. Damodar, and R. Edla, “Diagnosis of Coronavirus Disease From Chest X - Ray Images Using DenseNet - 169 Architecture,” *SN Comput. Sci.*, vol. 4, no. 3, pp. 1–6, 2023, doi: 10.1007/s42979-022-01627-7.
- [11] I. Sultan, H. Gharaibeh, A. Gharaibeh, B. Lahham, M. Al-tarawneh, and R. Al-qawabah, “LungVisionNet : A Hybrid Deep Learning Model for Chest X-Ray Classification — A Case Study at King Hussein Cancer Center (KHCC),” pp. 1–25, 2025.
- [12] C. Disease and D. Using, “A Review of Recent Advances in Deep Learning Models for Chest Disease Detection Using Radiography,” 2023, doi: doi.org/10.3390/diagnostics13010159.
- [13] S. Calderon-ramirez *et al.*, “Dealing with Scarce Labelled Data : Semi-supervised Deep Learning with Mix Match for Covid-19 Detection Using Chest X-ray Images,” pp. 5294–5301, 2021, doi: 10.1109/ICPR48806.2021.9412946.
- [14] A. R. Sajun and I. Zuolkernan, “applied sciences Investigating the Performance of FixMatch for COVID-19 Detection in Chest X-rays,” 2022, doi: doi.org/10.3390/app12094694.
- [15] A. H. Edoardo Vantaggiato, Emanuela Paladini, Fares Bougourzi, Cosimo Distanto, “COVID-19 Recognition Using Ensemble-CNNs in Two New Chest X-ray Databases,” pp. 1–20, 2021, doi: doi.org/10.3390/s21051742 Academic.
- [16] S. Zoha, J. G. Lee, and Y. W. Ko, “CAT: Class-aware adaptive-thresholding for robust semi-supervised domain generalization,” *PLoS One*, vol. 20, no. 9 September, pp. 1–20, 2025, doi: 10.1371/journal.pone.0329799.
- [17] D. Berthelot, R. Roelofs, K. Sohn, N. Carlini, and A. Kurakin, “Adamatch: a Unified Approach To Semi-Supervised Learning and Domain Adaptation,” *ICLR 2022 - 10th Int. Conf. Learn. Represent.*, no. 2, pp. 1–50, 2022.
- [18] U. Chutia, A. Shanker, T. Jyoti, P. Singh, and V. Kumar, “Classification of Lung Diseases Using an Attention - Based Modified DenseNet Model,” pp. 1625–1641, 2024, doi: 10.1007/s10278-024-01005-0.
- [19] S. Shastri, I. Kansal, S. Kumar, K. Singh, R. Popli, and V. Mansotra, “CheXImageNet : a novel architecture for accurate classification of Covid - 19 with chest x - ray digital images using deep convolutional neural networks,” *Health Technol. (Berl.)*, pp. 193–204, 2022, doi: 10.1007/s12553-021-00630-x.
- [20] S. Kumar and R. Rani, “LiteCovidNet : A lightweight deep neural network model for detection of COVID-19 using X-ray images,” no. March 2021, pp. 1464–1480, 2022, doi: 10.1002/ima.22770.
- [21] D. Zhang, F. Ren, Y. Li, L. Na, and Y. Ma, “Pneumonia Detection from Chest X-ray Images Based on Convolutional Neural Network,” 2021, doi: doi.org/10.3390/electronics10131512 1.
- [22] C. Mosquera, L. Ferrer, D. H. Milone, D. Luna, and E. Ferrante, “Class imbalance on medical image classification: towards better evaluation practices for discrimination and calibration performance.,” *Eur. Radiol.*, vol. 34, no. 12, pp. 7895–7903, Dec. 2024, doi: 10.1007/s00330-024-10834-0.
- [23] M. Elgendi *et al.*, “The Effectiveness of Image Augmentation in Deep Learning Networks for Detecting COVID-19: A Geometric Transformation Perspective,” vol. 8, no. March, pp. 1–12, 2021, doi: 10.3389/fmed.2021.629134.
- [24] D. L. Rubin, C. P. Langlotz, and A. S. Chaudhari, “Exploring Image Augmentations for Siamese Representation Learning with Chest X-Rays,” pp. 444–467, 2023.
- [25] A. Ke, W. Ellsworth, and A. Y. Ng, *CheXtransfer : Performance and Parameter Efficiency of ImageNet Models for Chest X-Ray Interpretation*, vol. 1, no. 1. Association for Computing Machinery, 2021. doi: 10.1145/3450439.3451867.
- [26] B. Kocak *et al.*, “European Journal of Radiology Artificial Intelligence Evaluation metrics in medical imaging AI : fundamentals , pitfalls , misapplications , and recommendations,” vol. 3, no. July, 2025, doi: 10.1016/j.ejrai.2025.100030.
- [27] R. Rs, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization,” *Int. J. Comput. Vis.*, vol. 128, Feb. 2020, doi: 10.1007/s11263-019-01228-7.