TIN: Terapan Informatika Nusantara
Vol 6, No 12, May 2026, page 2345-2360
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i12.9848

Klasifikasi Multikelas Citra Chest X-Ray Menggunakan Semi-Supervised SoftMatch pada Label Terbatas

M. Nabil Dawami, Benny Sukma Negara[*], Muhammad Irsyad, Yusra, Febi Yanto

Fakultas Sains dan Teknologi, Program Studi Teknik Informatika, Universitas Islam Negeri Sultan Syarif Kasim, Pekanbaru, Indonesia
Email: [1]12250111527@students.uin-suska.ac.id, [2,*]bsnegara@uin-suska.ac.id, [3]irsyadtech@uin-suska.ac.id, [4]yusra@uin-suska.ac.id, [5]febiyanto@uin-suska.ac.id
Email Penulis Korespondensi: bsnegara@uin-suska.ac.id

Abstrak–Klasifikasi citra *chest X-ray* (CXR) berbasis *deep learning* kerap menghadapi kendala berupa kelangkaan data medis berlabel dan distribusi kelas yang tidak seimbang. Penelitian ini bertujuan mengimplementasikan pendekatan *semi-supervised learning* (SSL) berbasis algoritma SoftMatch dengan *backbone* DenseNet-121 untuk klasifikasi multikelas citra CXR (Normal, COVID-19, dan Pneumonia) pada kondisi keterbatasan label. SoftMatch dipilih secara spesifik karena kemampuannya memitigasi *quantity-quality trade-off* melalui mekanisme *soft-weighting pseudo-label* secara adaptif. Dataset sebanyak 5.228 citra dialokasikan menggunakan *stratified split* menjadi 70% data latih, 10% data validasi, dan 20% data uji. Eksperimen dilakukan pada tiga skenario proporsi data berlabel, yaitu 5%, 10%, dan 20%, masing-masing dengan dan tanpa *Uniform Alignment*. Evaluasi dilakukan menggunakan metrik *accuracy*, *macro F1-score*, *confusion matrix*, ROC-AUC, serta didukung analisis interpretabilitas visual menggunakan Grad-CAM. Hasil pengujian menunjukkan bahwa model tetap tangguh pada skenario paling kritis (5% label) dengan mencapai *accuracy* 91,68% dan *macro F1-score* 91,72% saat mengintegrasikan *Uniform Alignment* (UA), mengungguli skenario tanpa UA yang mencatatkan *accuracy* 90,73% dan *macro F1-score* 90,82%. Performa terbaik konfigurasi UA diperoleh pada skenario 10% label (*accuracy* 94,46%; *macro F1-score* 94,58%), sedangkan performa puncak keseluruhan diraih oleh skenario 20% label tanpa UA (*accuracy* 95,79%; *macro F1-score* 95,89%). Temuan ini menunjukkan bahwa *Uniform Alignment* efektif pada kondisi label rendah hingga menengah, namun tidak selalu meningkatkan performa pada proporsi label yang lebih tinggi.

Kata Kunci: *Chest X-Ray*; *SoftMatch*; *Semi-Supervised Learning*; DenseNet-121; Klasifikasi Multikelas; *Uniform Aligment*

Abstract–Deep learning-based chest X-ray (CXR) classification frequently encounters bottlenecks due to the scarcity of labeled medical data and imbalanced class distributions. This study aims to implement a semi-supervised learning (SSL) approach utilizing the SoftMatch algorithm with a DenseNet-121 backbone for the multiclass classification of CXR images (Normal, COVID-19, and Pneumonia) under limited label conditions. SoftMatch is specifically selected for its capability to mitigate the quantity-quality trade-off through an adaptive pseudo-label soft-weighting mechanism. A dataset comprising 5,228 images is allocated via a stratified split into 70% training data, 10% validation data, and 20% testing data. Experiments are conducted across three labeled data proportion scenarios (5%, 10%, and 20%), each evaluated with and without Uniform Alignment. Evaluation metrics include accuracy, macro F1-score, confusion matrix, ROC-AUC, supported by visual interpretability analysis using Grad-CAM. The experimental results demonstrate that the model remains robust under the most critical scenario (5% labels), achieving an accuracy of 91.68% and a macro F1-score of 91.72% when integrating Uniform Alignment (UA), outperforming the scenario without UA, which records an accuracy of 90.73% and a macro F1-score of 90.82%. The best performance for the UA configuration is achieved in the 10% label scenario (accuracy 94.46%; macro F1-score 94.58%), while the peak overall performance is attained by the 20% label scenario without UA (accuracy 95.79%; macro F1-score 95.89%). These findings indicate that Uniform Alignment is effective in low-to-medium label conditions but does not consistently enhance performance at higher label proportions.

Keywords: Chest X-Ray; SoftMatch; Semi-Supervised Learning; DenseNet-121; Multiclass Classification; Uniform Aligment

1. PENDAHULUAN

Infeksi paru akut, khususnya Pneumonia dan COVID-19, menghadirkan tantangan diagnostik yang signifikan. Manifestasi visual keduanya sering kali saling tumpang-tindih (*overlapping*) pada hasil pencitraan, sehingga sering sulit dibedakan jika hanya melalui inspeksi visual standar. Dalam kondisi darurat atau fasilitas kesehatan dengan sumber daya terbatas, *chest X-ray* (CXR) menjadi modalitas pilihan utama karena waktu akuisisi yang singkat dan ketersediaan alat yang luas (Wang et al., 2021). Oleh karena itu, CXR memegang peran sentral tidak hanya sebagai alat triase klinis, tetapi juga sebagai sumber data penting untuk pengembangan sistem diagnosis otomatis berbasis *deep learning* guna membedakan kelas Normal, COVID-19, dan Pneumonia secara cepat dan lebih presisi.

Meskipun *deep learning* telah menunjukkan kinerja yang andal, klasifikasi multikelas citra CXR masih menghadapi tantangan mendasar berupa tingginya ketergantungan pada data berlabel berskala besar. Proses anotasi citra medis membutuhkan keahlian klinis spesifik, serta memakan waktu, dan menimbulkan biaya yang tidak sedikit. Di sisi lain, beban kerja radiolog yang terus meningkat serta tuntutan kecepatan dalam pembacaan citra untuk dilaporkan, sangat memungkinkan terjadi nya kelelahan kognitif. Kondisi ini pada gilirannya dapat meningkatkan risiko *perceptual error* (kesalahan persepsi) dan *missed findings* (temuan terlewat), terutama pada kasus radiologis yang samar (Alexander et al., 2022; Kaviani et al., 2022). Kesenjangan antara lonjakan volume data CXR harian dan keterbatasan tenaga ahli penanda (*annotator*) medis menegaskan perlunya model yang tidak sepenuhnya bergantung pada *supervised learning* murni (Liu et al., 2021). Tantangan ini diperparah oleh ketidakseimbangan ketersediaan label antar kelas, yang membuat model rentan mengalami bias terhadap kelas mayoritas dan mengabaikan kelas minoritas (Calderon-Ramirez et al., 2021; Huynh et al., 2022).

TIN: Terapan Informatika Nusantara
Vol 6, No 12, May 2026, page 2345-2360
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i12.9848

Berbagai penelitian terdahulu telah berupaya mengurangi keterbatasan label melalui pendekatan *semi-supervised learning* (SSL), namun masih menyisakan ruang perbaikan. Pendekatan *teacher-student* (Liu et al., 2021) dan *pseudo-labeling* dengan *consistency regularization* (Sajun et al., 2022) dilaporkan mampu meningkatkan akurasi awal dengan memanfaatkan data tidak berlabel. Namun, model-model ini rentan terhadap penurunan performa pada skenario label yang sangat terbatas. Penelitian lain oleh Sultan et al. (2025) berhasil mencapai klasifikasi CXR yang akurat, tetapi model tersebut masih beroperasi secara *supervised* murni yang menuntut ketersediaan anotasi penuh. Sementara itu, kajian oleh der Sluijs et al. (2024) menyoroti kompleksitas ekstraksi fitur CXR. Mereka menggarisbawahi bahwa modifikasi pada citra medis sangat rentan mendistorsi pola patologis aslinya. Berdasarkan kajian literatur tersebut dan hasil eksperimen pada penelitian ini, terlihat bahwa ketika proporsi data berlabel turun di bawah sekitar 10%, model cenderung mengalami penurunan sensitivitas (*recall*) pada kelas minoritas. Model juga kesulitan mempertahankan margin pemisah yang jelas antar kelas penyakit paru yang memiliki karakteristik radiologis saling berdekatan.

Salah satu sumber keterbatasan tersebut berkaitan dengan mekanisme seleksi *pseudo-label* yang terlalu kaku melalui *hard thresholding* (misalnya pada FixMatch), yang memunculkan paradoks *quantity-quality trade-off* (Chen et al., 2023). Pada domain citra natural seperti CIFAR atau ImageNet, objek umumnya memiliki batas semantik yang tegas sehingga *hard thresholding* relatif efektif. Sebaliknya, pada citra CXR, variasi patologis bersifat kontinu dan sangat halus. Perbedaan tipis pada gradasi *ground-glass opacity* atau konsolidasi sering kali memunculkan tingkat keyakinan (*confidence*) menengah dari model. Mengabaikan prediksi menengah ini melalui *hard thresholding* berpotensi membuang sampel-sampel transisi medis yang justru penting untuk membentuk representasi fitur penyakit yang lebih komprehensif.

Untuk menjembatani celah kritis tersebut, penelitian ini mengimplementasikan SoftMatch, sebuah algoritma SSL yang beralih ke mekanisme *soft weighting*. Mekanisme ini memberikan bobot secara proporsional pada sampel tidak berlabel berdasarkan tingkat keyakinan prediksinya, sehingga seluruh spektrum data medis dapat dimanfaatkan tanpa mengorbankan kualitas sinyal pembelajaran (Chen et al., 2023). Dalam implementasinya, SoftMatch diintegrasikan dengan *backbone* DenseNet-121. Pemilihan arsitektur ini didasarkan pada sinergi teoretis: konektivitas padat (*dense connectivity*) pada DenseNet terbukti membantu mempertahankan fitur *low-level* seperti tekstur halus paru di lapisan dalam model (Huang et al., 2022; Quiñonez-Baca et al., 2025). Hal ini mendukung kalibrasi probabilitas kelas yang lebih presisi, yang merupakan prasyarat penting agar mekanisme pembobotan *soft weighting* pada SoftMatch dapat bekerja secara optimal.

Penelitian ini bertujuan mengimplementasikan dan mengevaluasi pendekatan SSL SoftMatch dengan DenseNet-121 sebagai *backbone* untuk klasifikasi multikelas CXR pada kondisi label sangat terbatas (hingga 5%). Kontribusi utama penelitian ini secara fundamental membedakan implementasi SoftMatch pada domain medis dengan penerapannya di domain visi komputer umum, yaitu melalui modifikasi intensitas augmentasi citra. Jika SSL standar di domain visi komputer mengandalkan augmentasi ekstrem (seperti rotasi tajam atau *cutout*) yang berisiko merusak makna anatomis paru-paru, penelitian ini mengkalibrasi augmentasi secara khusus untuk menjaga fidelitas klinis (der Sluijs et al., 2024). Kebaruan ini dipadukan dengan pengujian komponen *Uniform Alignment* (UA) untuk mengompensasi bias distribusi kelas akibat ketidakseimbangan subset berlabel. Evaluasi performa dilakukan menggunakan multi-metrik yang peka terhadap bias kelas (*macro F1-score*, ROC-AUC) dan interpretasi visual Grad-CAM. Hasil penelitian ini diharapkan dapat mendukung perancangan sistem diagnostik medis yang tangguh di tengah krisis data berlabel tanpa mengorbankan integritas klinis.

2. METODOLOGI PENELITIAN

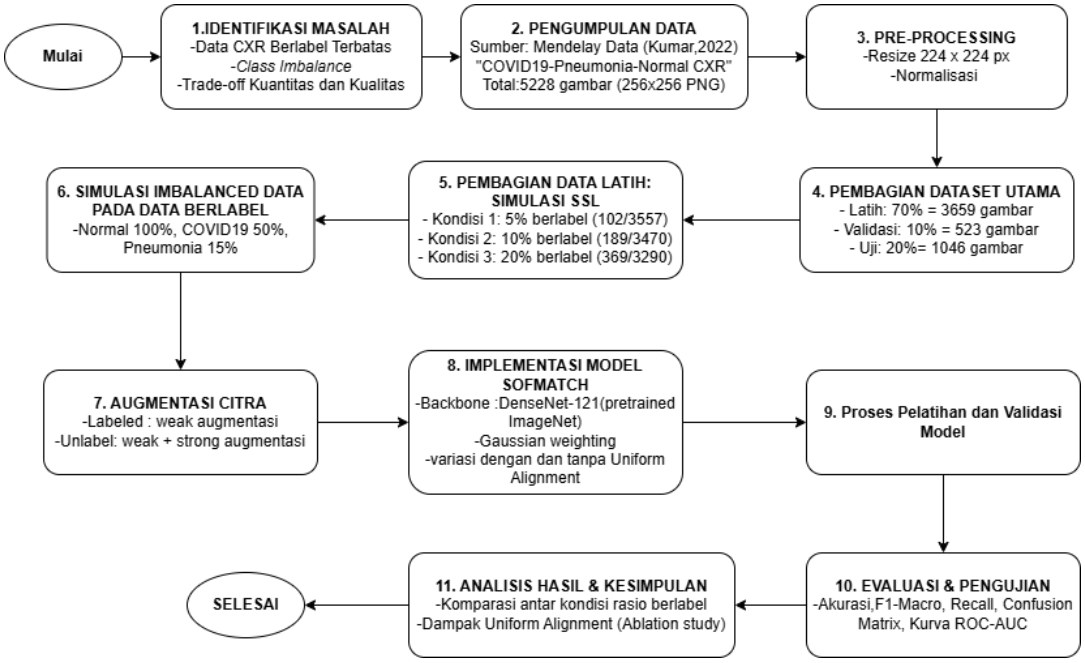
2.1 Desain Penelitian

Penelitian ini menggunakan pendekatan eksperimen komputasional dengan desain komparatif dan studi ablasi untuk mengevaluasi kinerja *semi-supervised deep learning* pada klasifikasi tiga kelas citra *chest X-ray*, yaitu Normal, COVID-19, dan Pneumonia. Model utama yang digunakan adalah SoftMatch dengan *backbone* DenseNet-121. Eksperimen dirancang untuk menilai dua aspek utama, yaitu pengaruh proporsi data berlabel dan kontribusi *Uniform Alignment* terhadap performa klasifikasi. Proporsi data berlabel ditetapkan dalam tiga skenario, yaitu 5%, 10%, dan 20%, sedangkan konfigurasi model dibedakan menjadi dua, yaitu SoftMatch dengan *Uniform Alignment* dan SoftMatch tanpa *Uniform Alignment*. Kombinasi tersebut menghasilkan enam skenario eksperimen yang diuji secara konsisten menggunakan dataset, *preprocessing*, *backbone*, parameter pelatihan, *validation set*, dan *test set* yang sama.

Sebagai pembanding pendukung, penelitian ini mengevaluasi *supervised baseline* DenseNet-121 secara spesifik pada skenario 5% dan 10% data berlabel. Pembatasan ini didasarkan pada landasan literatur bahwa titik kritis (*critical threshold*) kemerosotan drastis performa model *supervised* murni pada domain citra medis terjadi pada rasio label 10% ke bawah (Huynh et al., 2022; Sajun et al., 2022). Skenario 20% (lebih dari 700 citra berlabel dalam *train pool*) telah memasuki kategori ketersediaan label menengah (*moderate label regime*), di mana model *supervised* konvensional mulai mampu mencapai konvergensi yang relatif stabil. Oleh karena itu, evaluasi *baseline* murni difokuskan secara eksklusif pada skenario 5% dan 10% guna menyoroti secara tajam kontribusi SSL pada kondisi defisit data yang sesungguhnya (*extreme scarcity*), tanpa menggeser fokus utama ke pembahasan *semi-supervised learning*.

TIN: Terapan Informatika Nusantara
Vol 6, No 12, May 2026, page 2345-2360
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i12.9848

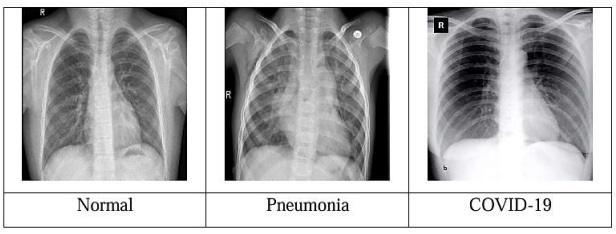
Alur penelitian secara sistematis divisualisasikan pada Gambar 1. Berdasarkan diagram tersebut, tahapan tidak disusun secara acak melainkan dirancang dengan keterkaitan logis yang ketat. Tahapan dimulai dari *preprocessing* untuk mengunci keseragaman distribusi nilai piksel input sebelum model dilatih. Dataset kemudian dipisah melalui *stratified split* untuk memastikan independensi data uji. Selanjutnya, simulasi ketidakseimbangan kelas secara sengaja hanya dikondisikan pada *train pool* agar ketahanan model SSL dapat diuji secara ekstrem, sementara *validation set* dan *test set* dipertahankan seimbang agar pengukuran performa tetap objektif. Skema augmentasi ganda kemudian diimplementasikan untuk menyediakan pasangan data (*weak-strong*) yang menjadi syarat utama berjalannya *consistency regularization* pada algoritma SoftMatch, sebelum akhirnya performa diskriminatif model diukur menggunakan multi-metrik.



Gambar 1. Flowchart Metodologi Penelitian

2.2 Pengumpulan Data

Dataset yang digunakan dalam penelitian ini adalah *Covid19-Pneumonia-Normal Chest X-Ray Images* dari repositori Mendeley Data (Kumar, 2022) yang berisikan 5.228 citra RGB serta berformat PNG (256 × 256 piksel). Secara klinis, dataset agregasi ini bersumber dari institusi medis berbeda dengan standar kurasi *ground truth* yang sangat ketat, kelas COVID-19 (1.626 citra) dikompilasi dari pasien dewasa pada repositori radiologi internasional seperti Eurorad dan Radiopaedia yang divalidasi langsung melalui konsensus dewan dokter spesialis radiologi, sedangkan kelas Normal (1.802 citra) dan Pneumonia (1.800 citra) diekstraksi dari kohort klinis pasien pediatrik (rentang usia 1–5 tahun) di *Guangzhou Women and Children's Medical Center* dengan label yang diverifikasi secara independen oleh dua orang dokter ahli. Distribusi komprehensif yang seimbang antar ketiga kelas ini tidak hanya memastikan keandalan medisnya, tetapi juga menjadikannya landasan dataset yang ideal dan bebas bias untuk skenario eksperimen *semi-supervised learning*.



Gambar 2. Contoh citra chest X-ray pada kelas Normal, COVID-19, dan Pneumonia

Pada penelitian ini, ketidakseimbangan kelas (*class imbalance*) tidak berasal dari dataset utama, melainkan diatur secara terkontrol khusus pada subset data berlabel di dalam *train pool*. Keputusan metodologis untuk menyimulasikan ketidakseimbangan ini *hanya* di dalam *train pool* didasarkan pada dua alasan esensial. Pertama, untuk mereplikasi kondisi pengembangan model klinis di dunia nyata, di mana volume data citra mentah sering kali melimpah, namun validasi anotasi spesialis untuk kelas penyakit tertentu sangat terbatas dan memakan waktu. Kedua, dengan membatasi ketidakseimbangan hanya pada tahap pelatihan, distribusi kelas pada *validation set* dan *test set* dapat dipertahankan tetap seimbang (melalui *stratified split*). Strategi ini sangat krusial untuk memastikan bahwa pengukuran

TIN: Terapan Informatika Nusantara
Vol 6, No 12, May 2026, page 2345-2360
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i12.9848

performa akhir model tetap objektif, tidak bias terhadap kelas mayoritas, dan secara adil merepresentasikan kemampuan model dalam mengenali setiap kelas. Dengan demikian, dataset utama tetap solid sebagai dasar eksperimen, sedangkan skenario simulasi digunakan untuk menguji batas ketahanan SoftMatch.

Pemilihan dataset ini juga didasarkan pada aspek keterbukaan akses dan pemenuhan prinsip *reproducibility*. Struktur data yang telah dikelompokkan secara terstruktur memudahkan penyusunan *pipeline* pelatihan, sedangkan statusnya sebagai repositori publik memungkinkan eksperimen ini direplikasi atau diverifikasi ulang oleh peneliti lain. Kelayakan dataset ini juga telah dibuktikan melalui penggunaannya secara luas pada berbagai literatur klasifikasi citra *chest X-ray* sebelumnya. Sebagai contoh, dataset ini digunakan dalam pengembangan arsitektur CheXImageNet oleh Shastri et al. (2022) serta model LiteCovidNet oleh Kumar et al. (2022). Hal tersebut mengonfirmasi bahwa dataset ini memiliki relevansi akademik yang kuat dan diakui secara ilmiah untuk penelitian diagnostik paru-paru berbasis kecerdasan buatan.

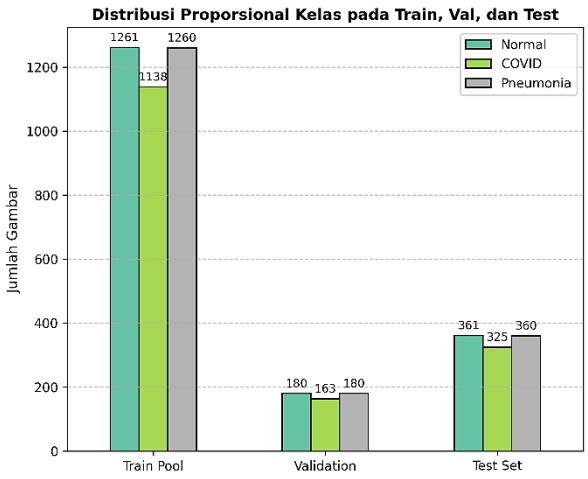
2.3 *Preprocessing dan Transformasi Input*

Tahap *preprocessing* dilakukan untuk menyeragamkan format citra sebelum digunakan pada pelatihan dan evaluasi model. Seluruh citra dibaca dalam format tiga *channel* RGB agar kompatibel dengan arsitektur DenseNet-121 yang menggunakan bobot *pretrained* ImageNet. Data latih diproses melalui jalur transformasi yang terintegrasi dengan skema augmentasi, sedangkan data validasi dan data uji menggunakan transformasi deterministik tanpa augmentasi agar evaluasi model tetap stabil. Pada data validasi dan uji, citra diubah langsung menjadi 224 × 224 piksel, kemudian dikonversi menjadi tensor.

Setelah penyesuaian ukuran, citra dinormalisasi menggunakan parameter statistik ImageNet, yaitu nilai *mean* = [0,485; 0,456; 0,406] dan *standard deviation* = [0,229; 0,224; 0,225]. Meskipun citra CXR secara inheren merupakan citra medis berskala abu-abu (*grayscale*) yang memiliki karakteristik visual berbeda dari citra natural, penerapan statistik ImageNet ini merupakan prosedur yang sangat krusial. Penyesuaian statistik ini tidak menghilangkan informasi diagnostik maupun menggeser distribusi rentang nilai penting pada paru-paru, karena ekstraksi fitur pada *deep learning* bergantung pada gradien spasial dan pola kontras relatif (seperti tepi tulang rusuk atau batas lesi konsolidasi), bukan pada pembacaan nilai intensitas absolut piksel tunggal. Lebih lanjut, karena model diinisialisasi menggunakan bobot *pretrained*, distribusi nilai piksel *input* baru wajib diselaraskan dengan distribusi data awal saat tahap *pretraining*. Keseragaman ini mencegah terjadinya pergeseran distribusi (*covariate shift*) yang dapat merusak bobot filter konvolusi pada lapisan-lapisan awal pengekstraksi tekstur (Ke et al., 2021). Oleh karena itu, strategi transformasi dan normalisasi konvensional ini digunakan untuk menjaga konsistensi representasi spasial, menjamin stabilitas proses *transfer learning*, dan telah tervalidasi secara empiris aman untuk klasifikasi paru-paru pada berbagai literatur medis mutakhir (Kamal et al., 2022; Sultan et al., 2025).

2.4 Pembagian Data

Dataset utama dibagi menjadi tiga subset menggunakan teknik *stratified split* dengan rasio 70% untuk *train pool*, 10% untuk *validation set*, dan 20% untuk *test set*. Teknik *stratified split* diterapkan untuk memastikan proporsi kelas Normal, COVID-19, dan Pneumonia tetap seimbang dan konsisten pada setiap subset data. Proses pembagian ini dikendalikan menggunakan nilai *random seed* sebesar 42 untuk menjamin aspek keterulangan (*reproducibility*) hasil partisi data. Berdasarkan proporsi tersebut, diperoleh alokasi sebanyak 3.659 citra pada *train pool*, 523 citra pada *validation set*, dan 1.046 citra pada *test set*.



Gambar 3. Distribusi kelas pada train pool, validation set, dan test set

Detail distribusi jumlah citra untuk masing-masing kelas pada ketiga subset tersebut disajikan secara visual pada Gambar 3. Pilihan metodologis menggunakan rasio pembagian data 70:10:20 ini mengadopsi standarisasi empiris yang lazim pada pengembangan *deep learning* untuk pencitraan medis. Alokasi ini secara optimal memisahkan fungsi

TIN: Terapan Informatika Nusantara
Vol 6, No 12, May 2026, page 2345-2360
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i12.9848

pelatihan, validasi hiperparameter, dan pengujian akhir secara independen, sehingga efektif dalam memitigasi risiko kebocoran data (*data leakage*) serta bias pembelajaran selama proses pemodelan berlangsung (Rajaraman et al., 2024; Zhang et al., 2022).

2.5 Simulasi Semi-Supervised Learning dan Ketidakseimbangan Subset Berlabel

Eksperimen *semi-supervised learning* dilakukan pada *train pool* dengan menetapkan proporsi ketersediaan data berlabel pada tiga skenario, yaitu 5%, 10%, dan 20%. Penentuan ambang batas bawah dan atas ini tidak dilakukan secara acak, melainkan didasarkan pada landasan teoretis performa *deep learning* medis. Angka 5% dipilih untuk merepresentasikan kondisi krisis ekstrem (*extreme scarcity*), di mana model *supervised* konvensional umumnya mengalami kegagalan konvergensi dan penurunan akurasi yang fatal (Huynh et al., 2022). Sebaliknya, angka 20% ditetapkan sebagai batas atas karena rasio ini merepresentasikan titik jenuh awal (*diminishing returns*) label menengah. Pada titik ini, penambahan ketersediaan data berlabel umumnya tidak lagi memberikan lonjakan performa SSL secara signifikan (Sajun et al., 2022) Rentang parameter ini ideal untuk menguji batas kemampuan, adaptabilitas, dan stabilitas algoritma SoftMatch secara komprehensif.

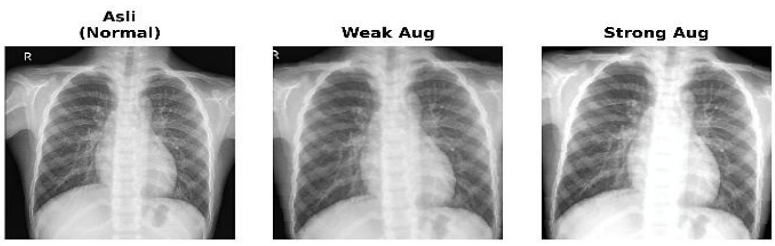
Pada tahap implementasi, kandidat data berlabel diekstraksi dari *train pool* menggunakan teknik *random stratified sampling* yang dikendalikan oleh *seed* tetap guna menjamin keterulangan (*reproducibility*) eksperimen. Setelah kandidat terbentuk, simulasi ketidakseimbangan direkayasa melalui mekanisme retensi kelas: Normal 100%, COVID-19 40%, dan Pneumonia 15%. Penetapan persentase hierarkis yang timpang ini merefleksikan realitas kesulitan anotasi klinis. Kelas Normal dipertahankan 100% karena ketersediaan citra paru sehat melimpah dan sangat mudah divalidasi. Kelas COVID-19 direpresi menjadi 40% untuk merepresentasikan kondisi triase pandemi, di mana kasus sangat tinggi namun kapasitas anotasi radiolog mulai kewalahan (Alexander et al., 2022). Sementara itu, retensi Pneumonia ditekan hingga 15% karena manifestasi visualnya sering kali tumpang-tindih, membutuhkan konsensus pakar tingkat lanjut yang memicu *bottleneck* (hambatan) pelabelan paling parah (Kaviani et al., 2022). Sampel yang tidak dipertahankan dialihkan ke kelompok data tidak berlabel agar tetap termanfaatkan. Melalui mekanisme ini, jumlah akhir data berlabel menjadi 102 citra (skenario 5%), 189 citra (10%), dan 369 citra (20%).

Penerapan kelas Pneumonia menjadi minoritas ekstrem (15%) mungkin tampak kontradiktif dengan keseimbangan dataset asli (sebagaimana dijelaskan pada Subbab 2.2). Namun, secara metodologis, hal ini sengaja dirancang sebagai sebuah *stress test* (uji ketahanan ekstrem). Dalam pengembangan kecerdasan buatan medis, volume akuisisi data mentah tidak selalu berbanding lurus dengan kecepatan ekstraksi label *ground truth*. Dengan memaksa model mengalami defisit label yang ekstrem pada kelas Pneumonia, eksperimen ini secara langsung menguji hipotesis utama penelitian: sejauh mana algoritma SoftMatch, khususnya komponen *Uniform Alignment* (UA), mampu memitigasi bias prediksi terhadap kelas mayoritas. Jika model mampu mempertahankan *recall* pada kelas minoritas Pneumonia di kondisi 15% ini, maka efektivitas kontribusi UA dalam mengatasi distorsi distribusi label dapat divalidasi secara kuat dan objektif (Chen et al., 2023).

2.6 Augmentasi Citra

Augmentasi citra diterapkan pada data pelatihan untuk mendukung *consistency regularization* dalam skema *semi-supervised learning*. Penelitian ini mengadopsi prinsip *weak-strong augmentation* guna membentuk *pseudo-label* yang stabil sekaligus menghitung konsistensi prediksi pada data tidak berlabel (Chen et al., 2023). Dalam penerapannya, domain citra medis menuntut pendekatan teoretis yang berbeda dari domain citra natural (seperti dataset CIFAR atau ImageNet). Objek pada citra natural umumnya bersifat invarian terhadap transformasi geometrik ekstrem. Sebaliknya, citra *chest X-ray* (CXR) memiliki batasan hierarki anatomis yang sangat kaku. Transformasi agresif pada CXR dapat menggeser posisi organ vital secara abnormal atau membuang indikator penyakit yang samar, sehingga model berisiko mempelajari representasi fitur yang cacat secara klinis (der Sluijs et al., 2024).

Oleh karena itu, konfigurasi augmentasi disesuaikan secara moderat agar relevan dengan karakteristik visual medis. Pada data berlabel dan jalur pertama data tidak berlabel, model menggunakan *weak augmentation* yang terdiri atas proses *resize* ke 256 × 256 piksel, *random rotation* 5°, *random affine translation* 0,02, dan *center crop* ke 224 × 224 piksel. Jalur kedua pada data tidak berlabel menantang model menggunakan *strong augmentation* yang terdiri atas *resize* ke 256 × 256 piksel, *random rotation* 10°, *color jitter* pada *brightness* dan *contrast* sebesar 0,3, *center crop* ke 224 × 224 piksel, dan RandAugment (2 operasi, magnitudo 5). Seluruh pemrosesan diakhiri dengan konversi ke tensor dan normalisasi.



Gambar 4. Ilustrasi weak augmentation dan strong augmentation pada citra chest X-ray

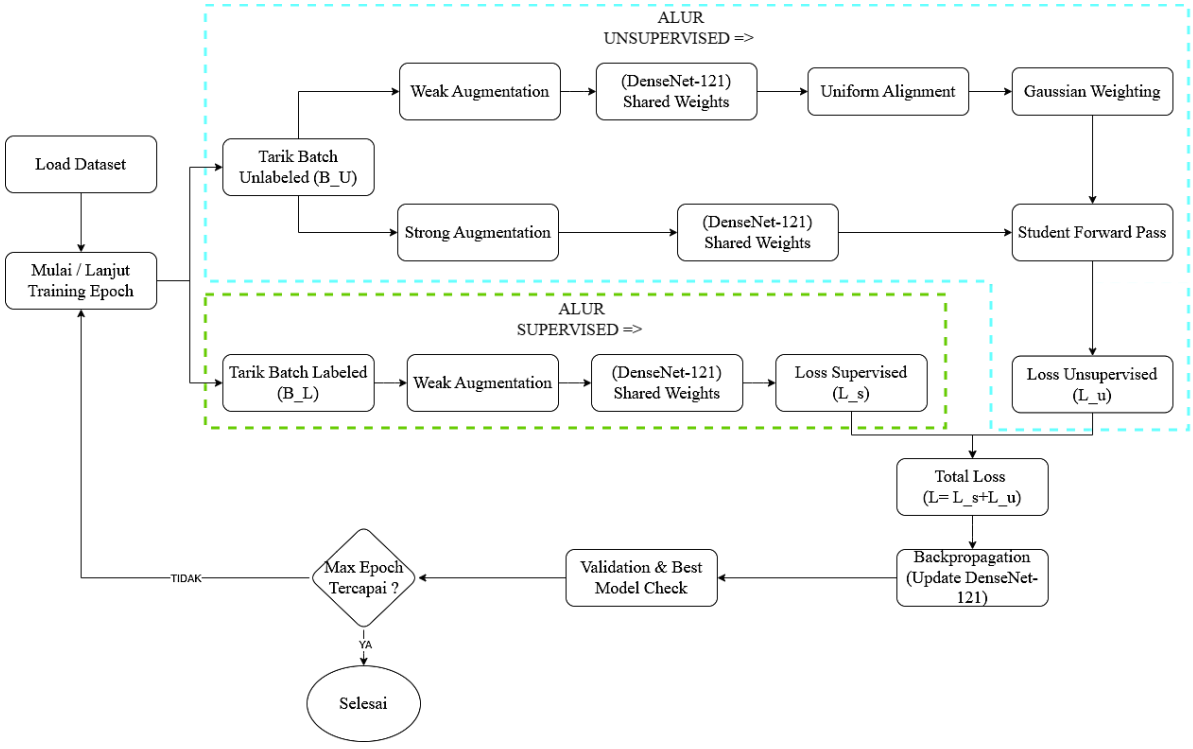
TIN: Terapan Informatika Nusantara
Vol 6, No 12, May 2026, page 2345-2360
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i12.9848

Contoh *output* transformasi disajikan pada Gambar 4. Penentuan ambang batas kuantitatif augmentasi tersebut didasarkan secara ketat pada toleransi radiologis. Sudut rotasi maksimal dikunci pada angka 10° karena rotasi di atas 10°–15° secara klinis terbukti mendistorsi evaluasi rasio kardiotorasik dan memunculkan artefak visual yang menyerupai kesalahan pemosisian (*malpositioning*) pasien saat akuisisi (Elgendi et al., 2021). Selain itu, modifikasi intensitas kecerahan dan kontras dibatasi maksimal pada skala 0,3. Manipulasi warna yang lebih tinggi dari batas tersebut akan memicu efek *overexposure*, yang berisiko menghapus gradasi halus lesi *ground-glass opacities* (GGO) pada COVID-19 maupun menyamarkan batas infiltrat konsolidasi pada Pneumonia (der Sluijs et al., 2024). Melalui pengaturan kuantitatif ini, keberagaman data berhasil ditingkatkan guna mendukung pembelajaran konsistensi tanpa merusak satupun informasi radiologis yang bermakna.

2.7 Implementasi Model SoftMatch dengan DenseNet-121

Model utama pada penelitian ini menggunakan DenseNet-121 sebagai backbone klasifikasi. Arsitektur ini dipilih karena telah terbukti relevan dalam klasifikasi citra chest X-ray, mendukung transfer learning, dan tetap kompetitif pada kondisi data berlabel terbatas. (Huang et al., 2022; Quiñonez-Baca et al., 2025; Wang et al., 2021) Pada implementasinya, DenseNet-121 diinisialisasi menggunakan bobot pretrained ImageNet, kemudian lapisan klasifikasi akhir diganti menjadi linear layer dengan tiga unit keluaran untuk merepresentasikan kelas Normal, COVID-19, dan Pneumonia. Penggunaan bobot pretrained dipertahankan karena transfer learning dari ImageNet dapat membantu stabilitas pelatihan ketika jumlah data berlabel tidak besar (Ke et al., 2021)

Implementasi *semi-supervised learning* dilakukan menggunakan *SoftMatch*. Pada setiap iterasi pelatihan, model menerima satu *batch* data berlabel dan satu *batch* data tidak berlabel. Data berlabel digunakan untuk menghitung *supervised loss*. Pada data tidak berlabel, prediksi dari citra hasil *weak augmentation* digunakan untuk membentuk *pseudo-label* dan nilai *confidence*, sedangkan citra hasil *strong augmentation* digunakan untuk menghitung *unsupervised loss*. Dengan skema ini, model tidak hanya belajar dari label aktual, tetapi juga memanfaatkan data tidak berlabel melalui pembobotan kontribusi *pseudo-label*. Pendekatan ini sesuai dengan karakteristik *SoftMatch* yang menggunakan *soft weighting* untuk mengurangi masalah *quantity-quality trade-off* pada *pseudo-labeling* (Chen et al., 2023). Alur implementasi secara umum ditunjukkan pada Gambar 5.



Gambar 5. Flowchart Alur Implementasi SoftMatch

Pada konfigurasi dengan *Uniform Alignment*, distribusi probabilitas prediksi pada data tidak berlabel diselaraskan terhadap distribusi target *uniform* sebelum *confidence* digunakan sebagai dasar pembobotan. Hasil *alignment* tidak digunakan langsung untuk membentuk *pseudo-label*, tetapi digunakan untuk menghitung bobot sampel dalam *unsupervised loss*. Setelah itu, *SoftMatch* menerapkan pembobotan berbasis *truncated Gaussian* agar sampel dengan *confidence* menengah tetap dapat dimanfaatkan secara proporsional. Pada konfigurasi tanpa *Uniform Alignment*, bobot dihitung langsung dari *confidence* prediksi awal. Perbandingan kedua konfigurasi ini digunakan sebagai studi ablasi untuk menilai kontribusi *Uniform Alignment* terhadap performa model pada kondisi label terbatas dan subset berlabel yang tidak seimbang (Chen et al., 2023). Prosedur / algoritma pelatihan *SoftMatch-DenseNet-121* diringkas sebagai berikut.

TIN: Terapan Informatika Nusantara
Vol 6, No 12, May 2026, page 2345-2360
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i12.9848

1. Masukkan data berlabel D_L, data tidak berlabel D_U, dan data validasi D_V, dengan jumlah kelas $C = 3$, yaitu *Normal*, *COVID-19*, dan *Pneumonia*.
2. Inisialisasi model DenseNet-121 menggunakan bobot *pretrained ImageNet*, kemudian ganti lapisan klasifikasi akhir menjadi *linear layer* dengan tiga unit keluaran.
3. Tentukan parameter pelatihan, yaitu jumlah *epoch* $T = 40$, *learning rate* $\eta = 0{,}003$, *momentum* $\beta = 0{,}9$, *weight decay* $\omega = 5 \times 10^{-4}$, ukuran *batch* data tidak berlabel $B_U = 48$, serta status penggunaan *Uniform Alignment*.
4. Tentukan ukuran *batch* data berlabel dari jumlah data *labeled* pada skenario ke-s, yaitu:

$$B_L = \max\left(4, \min\left(16, \left\lfloor \frac{N_L}{4} \right\rfloor\right)\right) \tag{1}$$

5. Inisialisasi parameter *Exponential Moving Average* untuk statistik *confidence* dan distribusi prediksi pada data tidak berlabel, yaitu:

$$\mu_t = \frac{1}{C}, \sigma_t^2 = 1, \hat{E}_{B_U}[p] = \left[\frac{1}{C}, \frac{1}{C}, \ldots, \frac{1}{C}\right] \tag{2}$$

6. Untuk setiap *epoch*, ambil satu *batch* data berlabel (x_l, y_l)berukuran $B_L$dan satu *batch* data tidak berlabel (u_w, u_s)berukuran B_U, dengan $u_w$merupakan citra hasil *weak augmentation* dan $u_s$merupakan citra hasil *strong augmentation*.
7. Hitung *supervised loss* pada data berlabel:

$$L_s = H(y_l, p(y \mid x_l)) \tag{3}$$

8. Hitung probabilitas prediksi pada data tidak berlabel hasil *weak augmentation*:

$$p = p(y \mid u_w) \tag{4}$$

9. Bentuk *pseudo-label* dan *confidence* awal dari prediksi cabang *weak*:

$$\hat{p} = \arg\max(p), c = \max(p) \tag{5}$$

10. Perbarui parameter *Exponential Moving Average* untuk rata-rata *confidence*, varians, dan distribusi prediksi *batch* tidak berlabel.
11. Jika *Uniform Alignment* diaktifkan, sesuaikan distribusi probabilitas prediksi terhadap distribusi target uniform:

$$UA(p) = Normalize\left(p \cdot \frac{u(C)}{\hat{E}_{B_U}[p]}\right) \tag{6}$$

12. Tentukan *confidence* untuk pembobotan:
    a. Jika menggunakan *Uniform Alignment*,

$$c_w = \max(UA(p)) \tag{7}$$

    b. Jika tidak menggunakan *Uniform Alignment*,

$$c_w = \max(p) \tag{8}$$

13. Hitung bobot sampel menggunakan fungsi *truncated Gaussian*:

$$\lambda(p) = \begin{cases} \exp\left(-\frac{(c_w - \mu_t)^2}{2\sigma_t^2}\right), & \text{jika } c_w < \mu_t \\ 1, & \text{lainnya} \end{cases} \tag{9}$$

14. Hitung *unsupervised loss* pada data tidak berlabel hasil *strong augmentation*:

$$L_u = \frac{1}{B_U} \sum_{i=1}^{B_U} \lambda(p_i)\, H(\hat{p}_i, p(y \mid u_{s,i})) \tag{10}$$

15. Hitung total *loss*:

$$L = L_s + L_u \tag{11}$$

16. Lakukan *backpropagation* dan pembaruan parameter model menggunakan *optimizer SGD*.
17. Pada akhir setiap *epoch*, evaluasi model pada data validasi. Jika *validation accuracy* pada *epoch* berjalan lebih baik dibandingkan model terbaik sebelumnya, maka simpan bobot model tersebut sebagai model terbaik.
18. Ulangi Langkah 6 sampai Langkah 17 hingga seluruh *epoch* selesai, kemudian gunakan model terbaik berdasarkan *validation accuracy* untuk evaluasi akhir pada data uji.

Eksperimen dijalankan pada enam konfigurasi, yaitu kombinasi skenario 5%, 10%, dan 20% data berlabel dengan dua kondisi model, yaitu menggunakan *Uniform Alignment* dan tanpa *Uniform Alignment*. Proses pelatihan dilakukan selama 40 *epoch* menggunakan *optimizer Stochastic Gradient Descent* (SGD) dengan *learning rate* awal 0,003, *momentum* 0,9, dan *weight decay* 5×10^{-4}. Penjadwalan *learning rate* dilakukan menggunakan

TIN: Terapan Informatika Nusantara
Vol 6, No 12, May 2026, page 2345-2360
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i12.9848

CosineAnnealingLR agar proses optimasi lebih stabil. *Batch* data tidak berlabel ditetapkan sebesar 48, sedangkan *batch* data berlabel dibuat dinamis sesuai jumlah data berlabel pada setiap skenario, dibatasi dengan nilai minimum 4 dan maksimum 16 sampel per iterasi.. Model terbaik dipilih berdasarkan *validation accuracy* karena *validation set* dipertahankan dari hasil *stratified split* dan tidak dilibatkan dalam simulasi ketidakseimbangan subset berlabel. Evaluasi akhir tetap dilaporkan menggunakan metrik *multi-kriteria* agar performa setiap kelas dapat dianalisis secara lebih menyeluruh.

2.8 Evaluasi dan Analisis Hasil

Evaluasi akhir dilakukan menggunakan *test set* independen yang sama sekali tidak dilibatkan dalam proses pelatihan, validasi, maupun seleksi model terbaik. Kinerja diskriminatif model diukur menggunakan instrumen multi-metrik yang meliputi *accuracy*, *macro precision*, *macro recall*, *macro F1-score*, *recall* spesifik per kelas, *confusion matrix*, dan ROC-AUC. Pendekatan evaluasi multi-metrik mutlak diperlukan karena pada klasifikasi pencitraan medis yang mengalami ketidakseimbangan kelas ekstrem, metrik *accuracy* tunggal sering kali menghasilkan *accuracy paradox* (Kocak et al., 2025). Kondisi tersebut memberikan gambaran performa yang terlampau optimistis dengan cara menutupi kegagalan model dalam mendeteksi kelas yang kurang terwakili. Oleh karena itu, evaluasi dikalibrasi menggunakan metrik yang lebih sensitif terhadap representasi setiap kelas secara adil (Mosquera et al., 2024).

Untuk mendeteksi potensi kegagalan model pada kelas minoritas, penelitian ini bersandar pada metrik *macro precision*, *macro recall*, dan *macro F1-score*. Dengan skema *macro-averaging*, kalkulasi performa dilakukan secara terpisah untuk setiap kelas (Normal, COVID-19, Pneumonia), lalu dirata-ratakan dengan bobot yang sepenuhnya setara (rasio 1:1:1) tanpa dipengaruhi oleh jumlah sampel aktual penyusunnya. Mekanisme ini memastikan bahwa jika model mengalami kegagalan deteksi pada kelas minoritas ekstrem (seperti Pneumonia yang hanya memiliki retensi label 15%), maka nilai keseluruhan *macro F1-score* akan anjlok secara drastis (Kocak et al., 2025). Selain metrik agregat, *confusion matrix* dan *recall* spesifik per kelas dianalisis untuk membedah arah distribusi kesalahan prediksi (*misclassification*).

Lebih lanjut, kapabilitas model dalam membedakan antar-kelas pada berbagai ambang batas (*threshold*) keputusan diukur menggunakan kurva ROC-AUC. Mengingat penelitian ini merupakan studi klasifikasi multikelas (3 kelas), penghitungan ROC-AUC diadaptasi secara eksplisit menggunakan skema *One-vs-Rest* (OvR). Melalui pendekatan OvR, model dievaluasi secara biner dengan membandingkan satu kelas spesifik melawan gabungan dua kelas lainnya secara bergantian, sehingga kemampuan pemisahan patologis (misalnya Pneumonia vs Bukan Pneumonia) dapat diobservasi secara spesifik dan objektif (Mosquera et al., 2024). Sebagai pendukung validitas, penelitian ini turut menyertakan metode Grad-CAM. Grad-CAM diimplementasikan bukan sebagai alat diagnosis klinis final, melainkan sebagai fungsi interpretabilitas komputasional (AI *Explainability*) untuk memverifikasi secara visual bahwa keputusan *semi-supervised* model benar-benar difokuskan pada area region citra paru yang relevan (Kamal et al., 2022). Secara keseluruhan, analisis hasil dititikberatkan pada komparasi performa antar-skenario proporsi data berlabel serta pembuktian empiris mengenai sejauh mana komponen *Uniform Alignment* mampu mengamankan model dari bias kelas.

3. HASIL DAN PEMBAHASAN

3.1 Rekapitulasi Hasil Pengujian

Pengujian dilakukan pada enam skenario eksperimen: 5%, 10%, dan 20% data berlabel, masing-masing dengan konfigurasi *Uniform Alignment* (UA) dan tanpa UA. Semua skenario dievaluasi menggunakan *test set* independen yang sama untuk menjaga konsistensi dan memungkinkan perbandingan yang objektif. Rekapitulasi performa model berdasarkan metrik *accuracy*, *macro precision*, *macro recall*, *macro F1-score*, dan *recall* spesifik kelas Pneumonia dirangkum secara lengkap pada Tabel 1.

Tabel 1. Rekapitulasi performa model pada seluruh skenario eksperimen

Skenario	Accuracy (%)	Precision Macro (%)	Recall Macro (%)	F1 Macro (%)	Recall Pneumonia (%)
5% Labeled + UA	91,68	92,79	91,91	91,72	78,06
5% Labeled tanpa UA	90,73	92,39	90,99	90,82	75,28
10% Labeled + UA	94,46	94,85	94,55	94,58	88,61
10% Labeled tanpa UA	94,17	94,96	94,32	94,31	84,72
20% Labeled + UA	93,31	93,81	93,40	93,37	85,28
20% Labeled tanpa UA	95,79	95,96	95,90	95,89	91,94

Berdasarkan Tabel 1, SoftMatch-DenseNet-121 berhasil mempertahankan performa tinggi di seluruh skenario, dengan tingkat *accuracy* secara konsisten berada di atas 90% dan *macro F1-score* bergerak pada rentang 90,82% hingga 95,89%. Kinerja ini membuktikan secara empiris bahwa model mampu mempelajari representasi fitur anatomis secara adaptif meskipun jumlah data berlabel sangat dibatasi. Secara keseluruhan, performa agregat tertinggi dicapai pada skenario 20% *labeled* tanpa UA, sedangkan konfigurasi yang menggunakan *Uniform Alignment* (UA) mencapai titik performa optimalnya pada skenario 10% *labeled*, khususnya dalam mendongkrak *recall* pada kelas minoritas ekstrem (Pneumonia). Tren data ini menegaskan temuan awal bahwa efektivitas *Uniform Alignment* sangat bergantung

TIN: Terapan Informatika Nusantara
Vol 6, No 12, May 2026, page 2345-2360
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i12.9848

pada konteks rasio label (*context-dependent*), di mana mekanisme ini memberikan kontribusi yang signifikan pada kelompok label rendah hingga menengah, tetapi tidak selalu menguntungkan seiring bertambahnya jumlah data berlabel yang riil.

Secara ringkas, evaluasi performa global ini mengungkap dua temuan utama: pertama, skenario 20% *labeled* tanpa UA memberikan hasil terbaik secara keseluruhan; kedua, skenario 10% *labeled* dengan UA merupakan titik optimal bagi modul penyeimbang tersebut. Temuan fundamental ini menjadi dasar pijakan untuk analisis kausalitas lanjutan mengenai pengaruh rasio data berlabel dan studi ablasi kontribusi *Uniform Alignment* pada subbab-subbab berikutnya.

3.2 Pengaruh Rasio Data Berlabel terhadap Performa SoftMatch

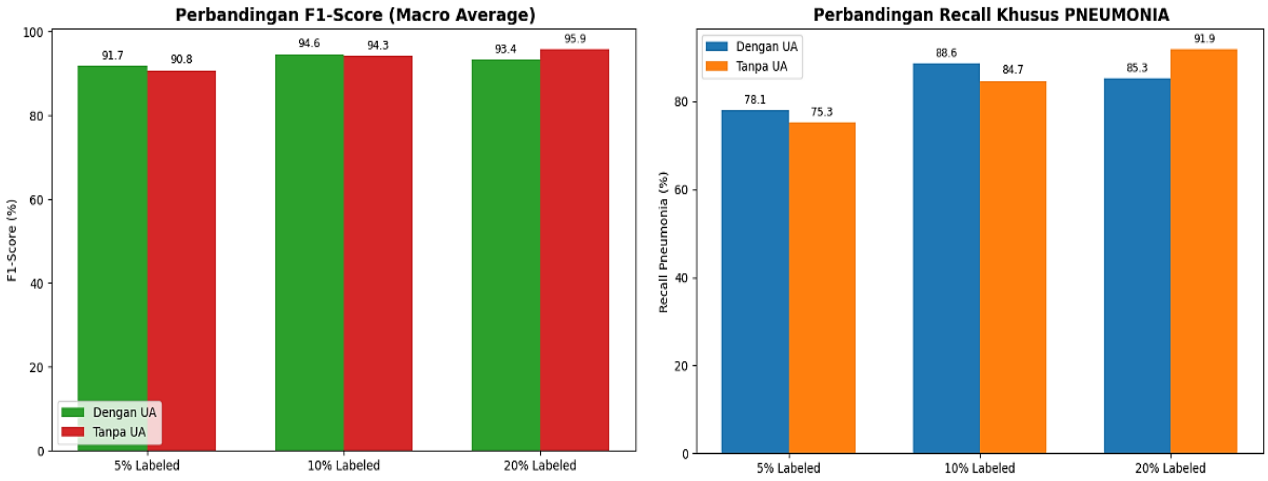
Pengaruh rasio data berlabel terhadap performa SoftMatch-DenseNet-121 menunjukkan pola tren yang berbeda secara signifikan antara konfigurasi dengan dan tanpa *Uniform Alignment* (UA). Pada konfigurasi tanpa UA, peningkatan proporsi data berlabel diikuti oleh peningkatan performa yang relatif linear, di mana nilai *macro F1-score* terus meningkat seiring bertambahnya jumlah label. Hal ini membuktikan secara empiris bahwa tambahan kuantitas data berlabel secara langsung memperkuat sinyal terawasi (*supervised signal*) dan membantu model memetakan representasi fitur kelas secara lebih efektif.

Sebaliknya, pada konfigurasi yang menggunakan UA, tren yang terbentuk bersifat non-linear. Kinerja optimal secara tak terduga dicapai pada skenario 10% *labeled*. Performa pada 5% *labeled* terpantau lebih rendah, dan ketika rasio data dinaikkan menjadi 20% *labeled*, performa model justru mengalami penurunan dibandingkan titik puncaknya di 10%. Pola ini memperlihatkan bahwa mekanisme UA sangat efektif beroperasi pada kondisi ketersediaan label yang terbatas atau kritis. UA mampu menyelaraskan distribusi *pseudo-label* untuk mengoreksi bias pembelajaran, namun intervensi penyeragaman ini tidak lagi memberikan keuntungan dan bahkan mulai mengganggu ketika sinyal terawasi (*supervised*) dari data berlabel sudah cukup dominan untuk memandu model secara mandiri.

Temuan ini merepresentasikan secara jelas karakteristik algoritma SoftMatch yang bertumpu pada pembobotan adaptif (*adaptive weighting*). Berdasarkan data eksperimen ini, terlihat jelas bahwa efektivitas *semi-supervised learning* pada klasifikasi citra medis sangat dipengaruhi oleh interaksi sensitif antara proporsi label aktual dan dinamika distribusi kelas. Dengan demikian, konfigurasi 10% *labeled* + UA terbukti menjadi titik ekuilibrium (keseimbangan) yang paling ideal antara pemanfaatan *pseudo-label* dan kekuatan sinyal *supervised*, sedangkan konfigurasi tanpa UA baru akan menunjukkan peningkatan performa yang lebih konsisten apabila kuantitas data berlabel dipastikan terus bertambah.

3.3 Studi Ablasi Uniform Alignment

Studi ablasi berfungsi untuk mengevaluasi kontribusi spesifik komponen *Uniform Alignment* (UA) dengan membandingkan metrik *macro F1-score* dan *recall* Pneumonia pada konfigurasi dengan dan tanpa UA, sebagaimana divisualisasikan pada Gambar 6.



Gambar 6. Perbandingan *macro F1-score* dan *recall Pneumonia* pada konfigurasi dengan UA dan tanpa UA

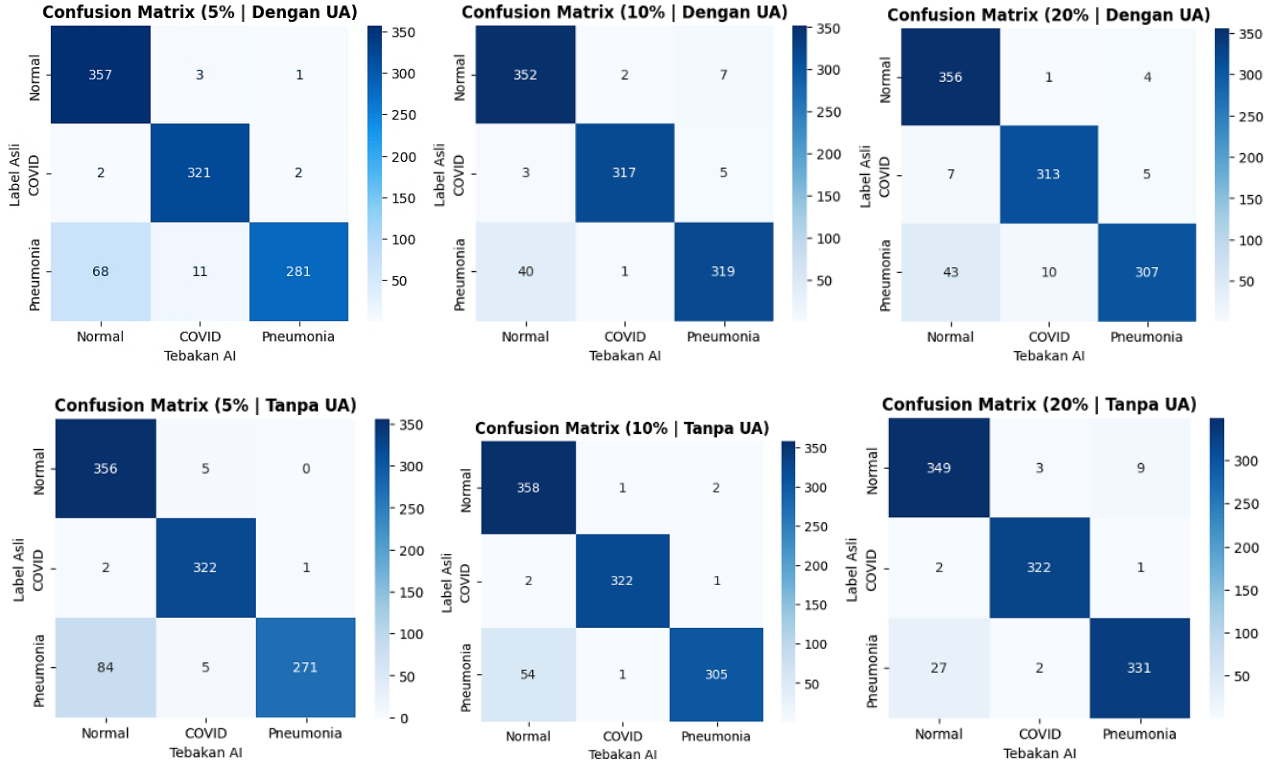
Berdasarkan Gambar 6, analisis kausalitas memperlihatkan peran vital UA pada kondisi kelangkaan label (5% dan 10%), di mana model rentan mengalami bias terhadap kelas dominan. Fungsi UA terbukti krusial dalam menekan bias tersebut dengan memaksa distribusi prediksi *pseudo-label* menjadi lebih seimbang (*uniform*). Secara kuantitatif, intervensi UA pada skenario 5% berhasil mendongkrak *recall* Pneumonia dari 75,28% menjadi 78,06%, serta menaikkan *macro F1-score* dari 90,82% menjadi 91,72%. Dampak penyelamatan paling optimal terjadi pada skenario 10%, di mana UA mencetak lonjakan signifikan pada *recall* Pneumonia dari 84,72% menjadi 88,61% (margin +3,89%) dengan *macro F1-score* puncak sebesar 94,58%.

TIN: Terapan Informatika Nusantara
Vol 6, No 12, May 2026, page 2345-2360
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i12.9848

Sebaliknya, pada skenario 20% data berlabel, fenomena berbalik drastis. Konfigurasi tanpa UA justru mendominasi dengan *macro F1-score* 95,89% dan *recall* Pneumonia 91,94%, jauh meninggalkan model dengan UA yang anjlok ke 93,37% dan 85,28% (mengalami defisit *recall* sebesar -6,66%). Secara analitis, anjloknya performa ini membuktikan bahwa pada kuantitas 20% label, model sejatinya telah menerima cukup informasi *supervised* untuk mempelajari dan memetakan batas keputusan (*decision boundary*) yang riil secara mandiri. Memaksakan koreksi distribusi menjadi *uniform* pada tahap ini justru mendistorsi tingkat keyakinan (*confidence*) model yang sudah akurat. Kesimpulannya, studi ablasi ini menegaskan bahwa UA adalah instrumen kontekstual: sangat efektif menekan bias pada defisit data ekstrem (≤10%), namun berbalik merugikan dan harus dinonaktifkan ketika kapasitas data *supervised* sudah mencukupi.

3.4 Analisis Per Kelas Berdasarkan Confusion Matrix

Pola sebaran prediksi model dianalisis secara mikroskopis melalui *confusion matrix* yang disajikan pada Gambar 7. Matriks tersebut secara konsisten menunjukkan bahwa perbedaan performa antar skenario sangat ditentukan oleh kemampuan model dalam mengklasifikasikan Pneumonia. Kelas Normal dan COVID-19 terbukti stabil dikenali, dengan rentang prediksi benar mencapai 349-358 dari 361 citra (Normal) dan 313-322 dari 325 citra (COVID-19). Sebaliknya, prediksi benar untuk Pneumonia sangat fluktuatif, bergerak lebar dari 271 hingga 331 dari total 360 citra, menegaskan bahwa kelas ini merupakan determinan utama keandalan model.



Gambar 7. Confusion matrix pada enam skenario eksperimen.

Pada konfigurasi dengan *Uniform Alignment* (UA), jenis kesalahan paling dominan adalah *false negative* (citra Pneumonia diprediksi sebagai Normal). Kesalahan ini berhasil ditekan dari 68 citra (5% *labeled* + UA) menjadi hanya 40 citra (10% *labeled* + UA), mengonfirmasi efektivitas UA pada kondisi label rendah hingga menengah. Namun, pada konfigurasi tanpa UA, kesalahan ini justru menyusut secara konsisten seiring bertambahnya proporsi data berlabel, mencapai titik terendah 27 citra pada skenario 20% *labeled* tanpa UA (menghasilkan 331 prediksi benar untuk Pneumonia). Hal ini membuktikan performa pengenalan kelas sulit yang paling stabil ketika sinyal *supervised* sudah mendominasi.

Secara keseluruhan, analisis matriks ini membuktikan secara empiris bahwa evaluasi model medis tidak dapat mengandalkan metrik agregat tunggal seperti *accuracy*. Keunggulan pendekatan SoftMatch-DenseNet-121 baru benar-benar tervalidasi melalui kemampuannya menyoroti dan menekan kesalahan kritis pada kelas Pneumonia. Model ini terbukti adaptif terhadap kelas yang paling sulit diidentifikasi, baik melalui bantuan kontekstual komponen UA saat kelangkaan label, maupun melalui pemanfaatan data mandiri saat label mencukupi.

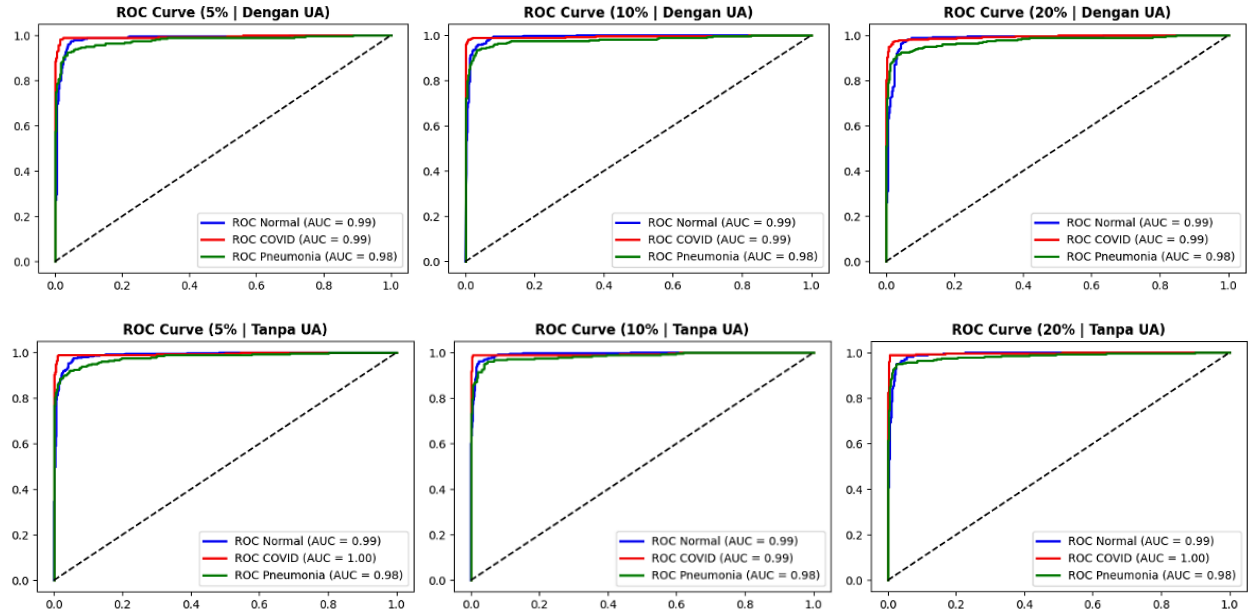
3.5 Analisis ROC Curve

Kemampuan diskriminatif probabilistik model SoftMatch-DenseNet-121 dalam memisahkan ketiga kelas citra *chest X-ray* pada berbagai ambang batas keputusan dievaluasi menggunakan pendekatan *one-vs-rest*, sebagaimana divisualisasikan pada Gambar 8. Berdasarkan grafik tersebut, seluruh kurva eksperimen melengkung tajam menjauhi

TIN: Terapan Informatika Nusantara
Vol 6, No 12, May 2026, page 2345-2360
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i12.9848

garis diagonal acak (*random classifier*). Nilai *Area Under Curve* (AUC) untuk pemisahan kelas Normal dan COVID-19 secara konsisten berada pada kisaran 0,99, sedangkan kelas Pneumonia berada pada kisaran 0,98. Capaian ini menandakan kemampuan probabilistik model yang sangat kuat dan efektif untuk membedakan tiap patologi secara individual.

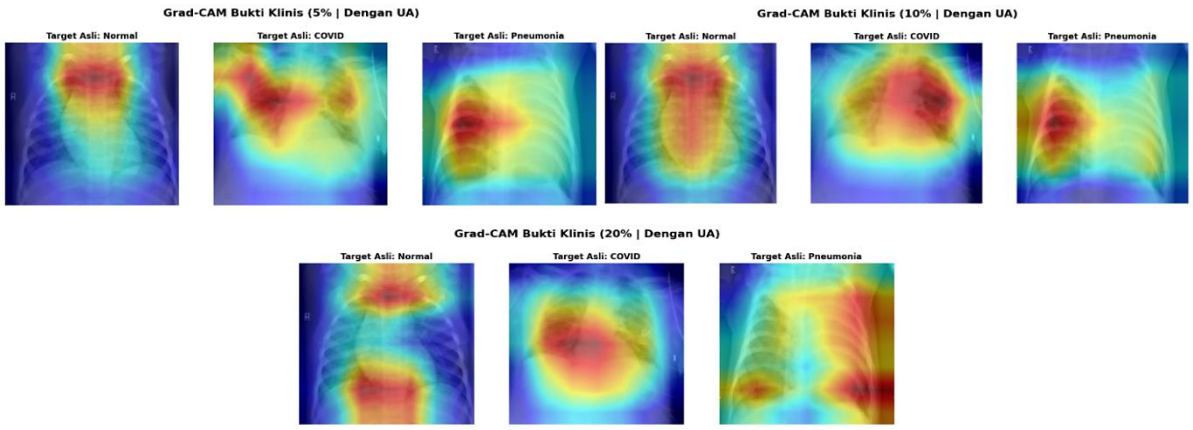
Meskipun memberikan nilai AUC agregat yang nyaris sempurna pada seluruh skenario, analisis ini membuktikan secara empiris bahwa metrik ROC-AUC kurang sensitif dalam menangkap celah performa akibat ketidakseimbangan subset berlabel. Perbedaan performa riil antar konfigurasi justru terekam jauh lebih jelas melalui metrik berbasis keputusan akhir, seperti *macro recall*, *macro F1-score*, *recall* Pneumonia, dan pola kesalahan prediksi pada *confusion matrix*. Temuan ini mempertegas bahwa pada klasifikasi medis dengan kelas minoritas yang fluktuatif, kurva ROC-AUC hanya ideal difungsikan sebagai indikator pemisah probabilitas global. Interpretasi akhir terhadap keandalan dan bias model mutlak harus dikonfrontasikan secara komprehensif menggunakan berbagai instrumen metrik kuantitatif.



Gambar 8. Kurva *ROC one-vs-rest* pada enam skenario eksperimen.

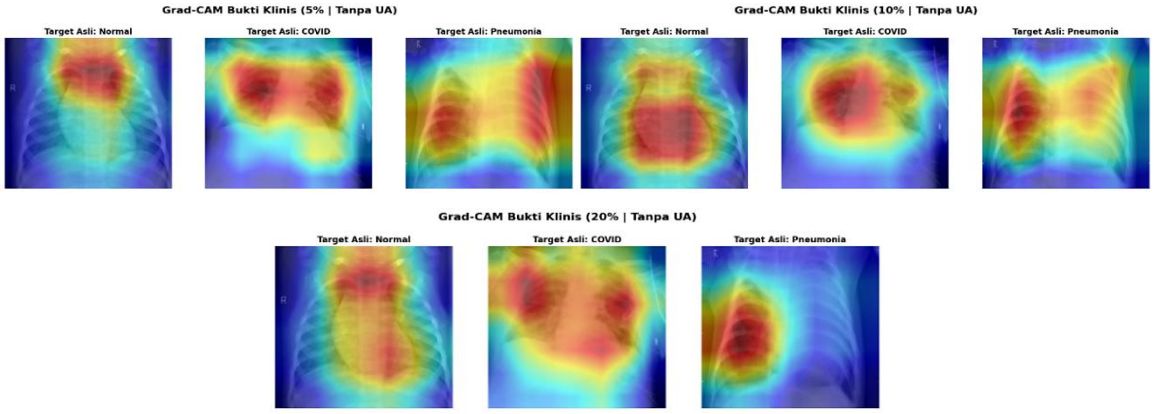
3.6 Analisis Interpretabilitas Visual Menggunakan Grad-CAM

Peta aktivasi visual (*heatmap*) diekstrak menggunakan metode Grad-CAM guna memverifikasi bahwa dasar keputusan *deep learning* sejalan dengan logika klinis medis. Analisis *computational interpretability* ini sangat penting untuk membuktikan bahwa model melakukan klasifikasi berdasarkan ekstraksi fitur patologis paru-paru riil, bukan akibat anomali komputasional seperti menghafal piksel artefak di luar paru (misalnya bayangan tulang rusuk, lengkung diafragma, atau teks label medis). Area berwarna panas (merah dan kuning) merepresentasikan wilayah *discriminative* dengan kontribusi tertinggi terhadap keputusan model, sedangkan area berwarna dingin (biru) menunjukkan wilayah yang diabaikan.



Gambar 9. Visualisasi Grad-CAM Menggunakan UA

TIN: Terapan Informatika Nusantara
Vol 6, No 12, May 2026, page 2345-2360
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i12.9848



Gambar 10. Visualisasi Grad-CAM Tanpa Menggunakan UA

Perbedaan representasi aktivasi untuk ketiga kelas serta pengaruh penggunaan *Uniform Alignment* (UA) dirangkum secara visual pada Gambar 9 dan Gambar 10. Berdasarkan visualisasi tersebut, model terbukti tidak melakukan kecurangan komputasional. Pada sampel uji kelas Normal, aktivasi model tampak menyebar secara merata untuk memantau kebersihan rongga lapangan paru (*lung fields*) tanpa memusatkan atensi pada anomali tertentu. Pada kasus COVID-19, model secara akurat memfokuskan piksel merah pada area periferal dan basal paru, yang secara medis merupakan lokasi klasik munculnya pola *ground-glass opacity* (GGO). Sementara itu, pada kasus Pneumonia, fokus atensi menembak tepat pada area opasitas pekat yang secara radiologis mengindikasikan keberadaan infiltrat konsolidasi lokal.

Lebih lanjut, perbandingan visual antara Gambar 9 (dengan UA) dan Gambar 10 (tanpa UA) mengungkap fungsi penyeimbang dari algoritma SoftMatch. Pada konfigurasi dengan UA di kondisi data sangat minim (10% *labeled*), fokus spasial terhadap lesi patologis terbukti lebih tajam dan terpusat (*localized*). Sebaliknya, pada konfigurasi tanpa UA, area warna panas cenderung lebih menyebar (*attention distribution* luas), meskipun metrik performa kuantitatifnya tetap tinggi pada skenario label yang lebih banyak. Secara keseluruhan, validasi visual ini mengonfirmasi bahwa intervensi *semi-supervised* SoftMatch-DenseNet-121 tidak merusak integritas representasi fitur medis; model secara komputasional mampu "melihat" dan melokalisasi penyakit paru layaknya prinsip diagnosis radiologis konvensional. Namun Penting untuk ditegaskan bahwa visualisasi Grad-CAM dalam penelitian ini murni berfungsi sebagai alat interpretabilitas komputasional (computational support). Meskipun atensi visual model terbukti sejalan dengan patologi medis, fokus perhatian ini tidak dapat ditafsirkan sebagai alat diagnosis atau validasi klinis final, melainkan sebagai instrumen untuk mendalami perilaku prediksi arsitektur SoftMatch-DenseNet-121.

3.7 Perbandingan dengan Model *Supervised Baseline*

Sebagai pembanding pendukung untuk menakar kontribusi objektif dari data tidak berlabel, penelitian ini mengevaluasi model *Supervised Baseline* DenseNet-121 yang murni hanya mengandalkan data berlabel (skenario 5% dan 10%). Perbandingan ini ditujukan untuk memberikan acuan performa *supervised* konvensional melawan pendekatan SoftMatch + *Uniform Alignment* (UA).

Tabel 2. *Supervised baseline dan SoftMatch + UA* pada skenario 5% dan 10% data berlabel

Skenario	Model	*Accuracy (%)*	*Precision Macro (%)*	*Recall Macro (%)*	*F1 Macro (%)*	*Recall Pneumonia(%)*
5% *Labeled*	DenseNet-121 *Baseline*	93,79	94,07	93,95	93,79	85,28
5% *Labeled*	*Sofmatch + UA*	91,68	92,79	91,91	91,72	78,06
10% *Labeled*	DenseNet-121 *Baseline*	94,65	94,83	94,77	94,74	89,72
10% *Labeled*	*Sofmatch + UA*	94,46	94,85	94,55	94,58	88,61

Berdasarkan Tabel 2, model *baseline* secara sekilas menampilkan performa agregat yang sangat tinggi, mencapai *macro F1-score* 93,79% pada 5% *labeled* dan 94,74% pada 10% *labeled*. Angka ini sedikit memimpin atau setara dengan capaian SoftMatch + UA (91,72% dan 94,58%). Meskipun capaian numerik *baseline* terlihat superior, evaluasi model medis tidak dapat bersandar secara eksklusif pada metrik akhir semata. Pencapaian tinggi dari arsitektur *deep learning* sekelas DenseNet-121 pada subset berlabel yang sangat ekstrem (hanya 5%) merupakan indikasi kuat dari fenomena penghafalan data latih (*memorization*).

Jika dikonfrontasikan dengan arsitektur terdahulu berbasis dataset identik, seperti model *CheXImageNet* atau *LiteCovidNet*, performa di atas 94% umumnya hanya bisa dicapai secara wajar jika model dilatih menggunakan 100% data penuh. Oleh karena itu, metrik tinggi *baseline* pada 5% label ini sangat berisiko merupakan ilusi statistik. Untuk membuktikan bahwa mekanisme *pseudo-labeling* dan *soft weighting* pada SoftMatch menawarkan keandalan (*robustness*) yang lebih superior daripada sekadar ilusi akurasi *baseline*, analisis konfrontasi harus digeser dari sekadar

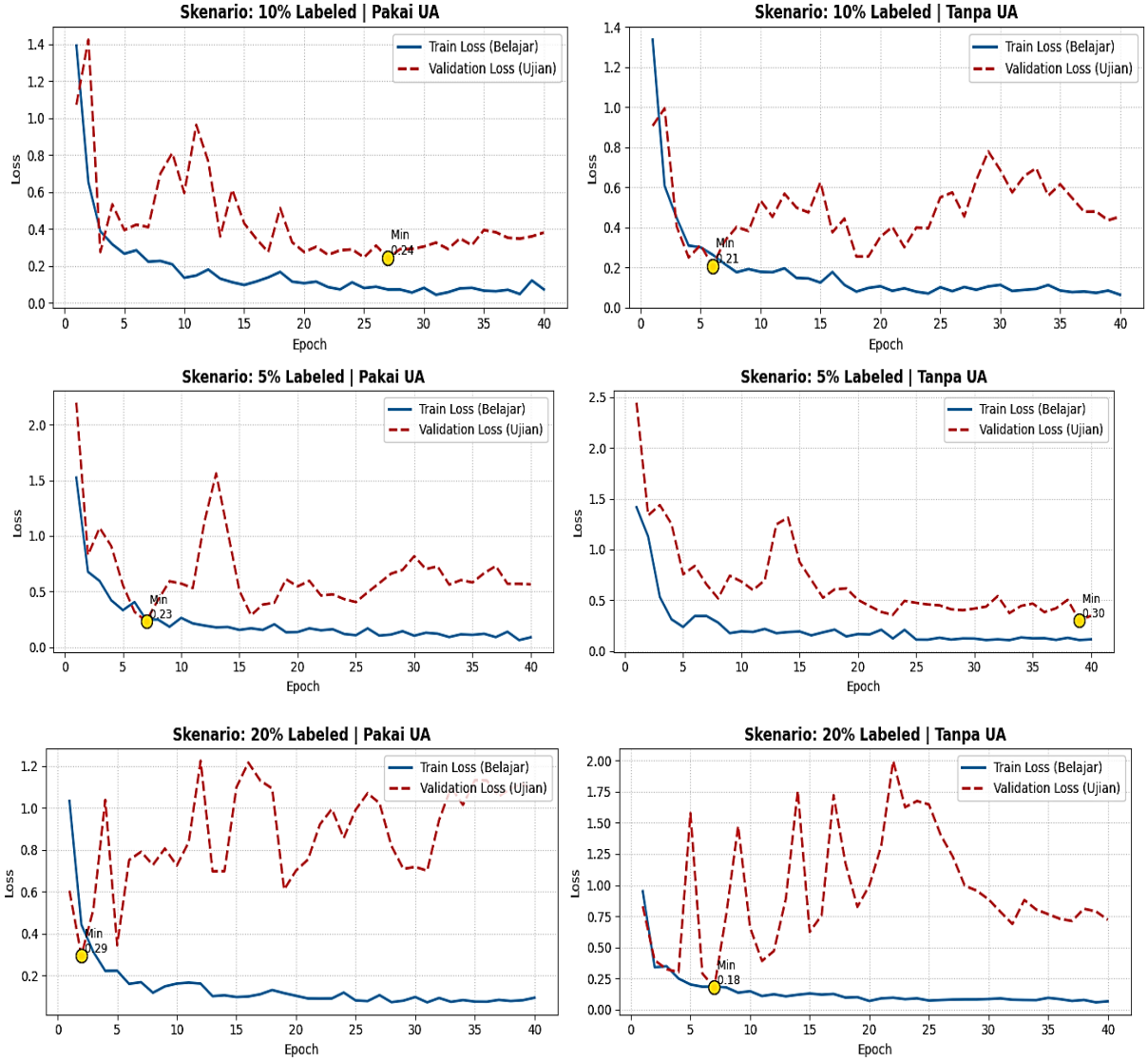
TIN: Terapan Informatika Nusantara
Vol 6, No 12, May 2026, page 2345-2360
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i12.9848

angka metrik akhir menuju stabilitas dinamika pembelajaran selama proses pelatihan (*generalization gap*), yang dibahas secara mendalam pada Subbab 3.8.

3.8 Analisis Dinamika Pembelajaran

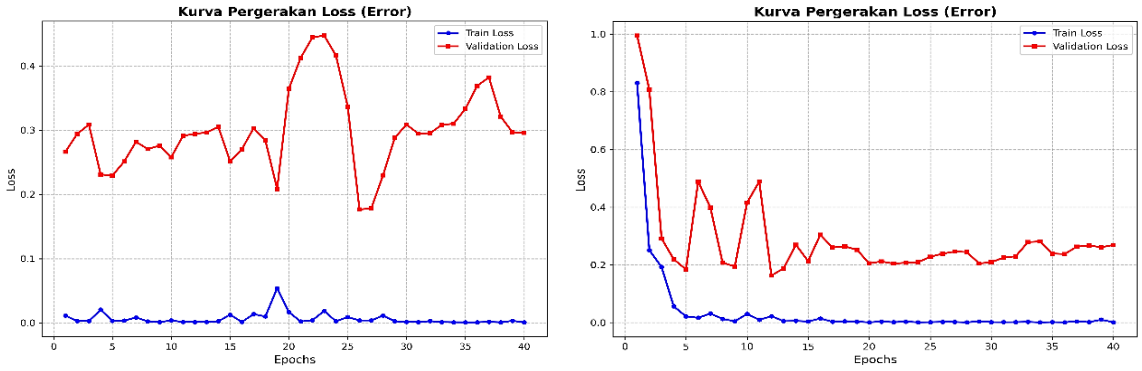
Kestabilan dan adaptabilitas model selama fase pelatihan dievaluasi melalui pelacakan tren *train loss* dan *validation loss*. Dinamika pembelajaran pada keenam skenario eksperimen SoftMatch divisualisasikan pada Gambar 11. Secara umum, lintasan *train loss* menurun konsisten pada seluruh skenario, sedangkan *validation loss* menunjukkan tingkat fluktuasi yang bervariasi bergantung pada ketersediaan label. Pada skenario 5% *labeled* (baik dengan maupun tanpa UA), tingginya *validation loss* merefleksikan krisis defisit sinyal *supervised*. Namun, konvergensi paling stabil dan mulus tercapai pada skenario 10% *labeled* + UA, selaras dengan capaian titik performa optimalnya. Sebaliknya, pada skenario 20%, *validation loss* yang paling konsisten menurun justru terjadi pada konfigurasi tanpa UA. Pola ini menegaskan kembali bahwa penyesuaian distribusi *pseudo-label* sangat adaptif di kondisi label terbatas, namun tidak lagi diperlukan saat sinyal *supervised* sudah mampu mendominasi arah pembelajaran.

Sebagai pembanding analitis, dinamika kelemahan model *Supervised Baseline* dibongkar secara gamblang melalui Gambar 12. Pada skenario 5%, *baseline* memperlihatkan fenomena *generalization gap* yang sangat parah, *train loss* anjlok drastis mendekati nol absolut, namun *validation loss* tertinggal jauh di atas dan sangat berfluktuasi. Fakta ini membuktikan bahwa capaian numerik metrik yang tinggi dari *baseline* sesungguhnya merupakan ilusi statistik akibat penyesuaian berlebihan terhadap data latih yang sangat sedikit (*overfitting* ekstrem). Meskipun dinamika *baseline* pada skenario 10% sedikit lebih terkendali, jarak (*gap*) pelebaran antara lintasan pelatihan dan validasi masih terlihat jelas dan belum sepenuhnya menghilangkan indikasi *overfitting*.



Gambar 11. Perbandingan train loss dan validation loss pada seluruh skenario eksperimen.

TIN: Terapan Informatika Nusantara
Vol 6, No 12, May 2026, page 2345-2360
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i12.9848



Gambar 12. Kurva loss supervised baseline pada skenario 5% dan 10% data berlabel.

Analisis dinamika *loss* ini mengonfirmasi keunggulan fundamental algoritma SoftMatch. Model *supervised baseline* terbukti hanya "kuat di atas kertas", tetapi rapuh dan gagal melakukan generalisasi yang aman secara klinis. Sebaliknya, pendekatan SoftMatch-DenseNet-121 memanfaatkan kelimpahan data tidak berlabel melalui *pseudo-label weighting* sebagai bentuk regularisasi yang tangguh. Hasilnya, model tidak hanya mencapai akurasi agregat yang tinggi, tetapi juga menunjukkannya melalui kurva konvergensi yang sehat dan stabil. Evaluasi ini memastikan bahwa keandalan diagnosis model AI medis tidak boleh hanya didasarkan pada metrik performa akhir, melainkan harus divalidasi melalui kestabilan dinamika pembelajarannya agar aman untuk implementasi dunia nyata.

3.9 Pembahasan

Hasil eksperimen menunjukkan bahwa implementasi SoftMatch-DenseNet-121 memberikan performa yang stabil pada kondisi data berlabel yang sangat terbatas. Berbeda dengan model *supervised baseline* yang terindikasi mengalami *overfitting* dan pelebaran *generalization gap* pada kurva pelatihannya (Gambar 12), kerangka kerja SoftMatch menunjukkan kemampuan dalam mempertahankan generalisasi model yang lebih baik melalui mekanisme *pseudo-labeling* dan pembobotan adaptif.

Secara komparatif di ranah *Semi-Supervised Learning* (SSL) medis, pendekatan SoftMatch menunjukkan potensi dalam mengurangi keterbatasan metode terdahulu. Kajian Sajun et al. (2022) melaporkan bahwa algoritma FixMatch yang menggunakan *hard thresholding* berpotensi mengabaikan sampel transisi, sehingga performa *macro F1-score* mereka menurun hingga 68% saat sampel label dibatasi secara ketat (20 citra per kelas). Sebaliknya, mekanisme *soft-weighting* pada penelitian ini memanfaatkan informasi dari sampel transisi secara lebih adaptif, dengan mencatatkan *macro F1-score* 91,72% dan *accuracy* 91,68% pada defisit label ekstrem (5% label atau setara ~34 citra per kelas).

Lebih lanjut, kerentanan model SSL konvensional terhadap bias kelas akibat *imbalance* (Calderon-Ramirez et al., 2021) berhasil dikurangi melalui mekanisme *Uniform Alignment* (UA). Intervensi penyeimbang UA menunjukkan peningkatan *recall* kelas minoritas ekstrem (Pneumonia) sebesar +2,78% pada skenario 5% label, dan kembali bertambah sebesar +3,89% (mencapai *recall* 88,61%) pada kondisi 10% label.

Jika dikonfrontasikan dengan model *supervised* murni, model ini menunjukkan efisiensi anotasi yang lebih tinggi. Model termutakhir seperti LungVisionNet (Sultan et al., 2025) mencetak akurasi 96,91%, namun hal tersebut menggunakan keseluruhan data berlabel (100% anotasi), serupa dengan persyaratan pada model CheXImageNet (Shastri et al., 2022) dan LiteCovidNet (Kumar et al., 2022) yang menggunakan dataset identik. Sebaliknya, pendekatan SoftMatch pada penelitian ini mampu mengurangi kebutuhan anotasi hingga 80% (skenario 20% label) namun tetap menyentuh performa kompetitif di angka *accuracy* 95,79%.

Meskipun demikian, data juga mengonfirmasi bahwa efektivitas modul UA bersifat kontekstual. Saat kuantitas label mencapai 20%, intervensi UA memicu defisit *recall* Pneumonia sebesar -6,66%, yang mengindikasikan bahwa penggunaan UA mungkin kurang optimal ketika data berlabel aktual sudah cukup dominan. Secara keseluruhan, validasi silang antara performa *macro F1-score* yang stabil dan interpretasi visual melalui Grad-CAM mengindikasikan bahwa model ini memiliki potensi untuk menekan ketergantungan label secara masif tanpa menunjukkan degradasi signifikan pada representasi patologis, sekaligus mendukung pentingnya penggunaan pendekatan multi-metrik dibandingkan hanya mengandalkan *accuracy* tunggal pada studi medis (Kocak et al., 2025).

4. KESIMPULAN

Penelitian ini menunjukkan bahwa pendekatan *semi-supervised deep learning* berbasis *SoftMatch* dengan DenseNet-121 mampu menghasilkan performa klasifikasi yang tergolong tinggi pada tiga kelas citra *chest X-ray*, yaitu Normal, COVID-19, dan Pneumonia, dalam kondisi data berlabel terbatas dan *subset* berlabel yang tidak seimbang. Pada konfigurasi dengan *Uniform Alignment*, performa terbaik diperoleh pada skenario 10% data berlabel dengan *accuracy* 94,46% dan *macro F1-score* 94,58%, sedangkan performa terbaik secara keseluruhan diperoleh pada skenario 20% data berlabel tanpa *Uniform Alignment* dengan *accuracy* 95,79% dan *macro F1-score* 95,89%. Temuan ini menunjukkan bahwa *Uniform Alignment* lebih efektif pada proporsi label rendah hingga menengah, tetapi tidak selalu meningkatkan

TIN: Terapan Informatika Nusantara
Vol 6, No 12, May 2026, page 2345-2360
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i12.9848

performa ketika jumlah data berlabel lebih besar karena sinyal *supervised* dari label aktual sudah lebih kuat. Perbandingan dengan *supervised baseline* menunjukkan bahwa DenseNet-121 mampu menghasilkan performa evaluasi yang tinggi ketika hanya menggunakan data berlabel, tetapi kurva *loss baseline* masih memperlihatkan *generalization gap* pada kondisi label terbatas. Hal ini memperkuat relevansi *SoftMatch* sebagai pendekatan *semi-supervised learning* karena model dapat memanfaatkan data tidak berlabel melalui *pseudo-labeling* dan pembobotan adaptif. Analisis per kelas menunjukkan bahwa Pneumonia menjadi kelas paling menantang karena lebih sering salah diklasifikasikan sebagai Normal, sehingga evaluasi tidak cukup hanya menggunakan *accuracy*, tetapi perlu mempertimbangkan *macro recall*, *macro F1-score*, *recall per kelas*, dan *confusion matrix*. Visualisasi *Grad-CAM* menunjukkan bahwa sebagian besar aktivasi model berada pada *region* toraks dan lapangan paru, tetapi tetap berperan sebagai interpretabilitas komputasional, bukan validasi klinis. Penelitian ini masih terbatas pada satu *dataset*, tiga kelas CXR paru-paru, dan satu *backbone* utama. Penelitian selanjutnya disarankan menggunakan *dataset* multi-sumber, mengeksplorasi *backbone* lain, membandingkan *SoftMatch* dengan metode *semi-supervised learning* lain, serta melibatkan validasi radiolog agar model lebih andal dalam konteks medis.

REFERENCES

Alexander, R., Waite, S., Bruno, M. A., Krupinski, E. A., Berlin, L., Macknik, S., & Martinez-Conde, S. (2022). Mandating Limits on Workload, Duty, and Speed in Radiology. *Radiology*, *304*(2), 274–282. https://doi.org/10.1148/radiol.212631

Calderon-Ramirez, S., Yang, S., Moemeni, A., Elizondo, D., Colreavy-Donnelly, S., Chavarría-Estrada, L. F., & Molina-Cabello, M. A. (2021). Correcting data imbalance for semi-supervised COVID-19 detection using X-ray chest images. *Applied Soft Computing*, *111*, 107692. https://doi.org/10.1016/j.asoc.2021.107692

Chen, H., Tao, R., Fan, Y., Wang, Y., Wang, J., Schiele, B., Xie, X., Raj, B., & Savvides, M. (2023). SoftMatch: Addressing the Quantity-Quality Tradeoff in Semi-supervised Learning. *The Eleventh International Conference on Learning Representations* . https://openreview.net/forum?id=ymt1zQXBDiF

der Sluijs, R. Van, Bhaskhar, N., Rubin, D., Langlotz, C., & Chaudhari, A. S. (2024). Exploring Image Augmentations for Siamese Representation Learning with Chest X-Rays. In I. Oguz, J. Noble, X. Li, M. Styner, C. Baumgartner, M. Rusu, T. Heinmann, D. Kontos, B. Landman, & B. Dawant (Eds.), *Medical Imaging with Deep Learning* (Vol. 227, pp. 444–467). PMLR. https://proceedings.mlr.press/v227/sluijs24a.html

Elgendi, M., Nasir, M. U., Tang, Q., Smith, D., Grenier, J.-P., Batte, C., Spieler, B., Leslie, W. D., Menon, C., Fletcher, R. R., Howard, N., Ward, R., Parker, W., & Nicolaou, S. (2021). The Effectiveness of Image Augmentation in Deep Learning Networks for Detecting COVID-19: A Geometric Transformation Perspective. *Frontiers in Medicine*, *8*. https://doi.org/10.3389/fmed.2021.629134

Huang, G.-H., Fu, Q.-J., Gu, M.-Z., Lu, N.-H., Liu, K.-Y., & Chen, T.-B. (2022). Deep Transfer Learning for the Multilabel Classification of Chest X-ray Images. *Diagnostics*, *12*(6). https://doi.org/10.3390/diagnostics12061457

Huynh, T., Nibali, A., & He, Z. (2022). Semi-supervised learning for medical image classification using imbalanced training data. *Computer Methods and Programs in Biomedicine*, *216*, 106628. https://doi.org/10.1016/j.cmpb.2022.106628

Kamal, U., Zunaed, M., Nizam, N. B., & Hasan, T. (2022). Anatomy-XNet: An Anatomy Aware Convolutional Neural Network for Thoracic Disease Classification in Chest X-Rays. *IEEE Journal of Biomedical and Health Informatics*, *26*(11), 5518–5528. https://doi.org/10.1109/JBHI.2022.3199594

Kaviani, P., Kalra, M. K., Digumarthy, S. R., Gupta, R. V., Dasegowda, G., Jagirdar, A., Gupta, S., Putha, P., Mahajan, V., Reddy, B., Venugopal, V. K., Tadepalli, M., Bizzo, B. C., & Dreyer, K. J. (2022). Frequency of Missed Findings on Chest Radiographs (CXRs) in an International, Multicenter Study: Application of AI to Reduce Missed Findings. *Diagnostics*, *12*(10), 2382. https://doi.org/10.3390/diagnostics12102382

Ke, A., Ellsworth, W., Banerjee, O., Ng, A. Y., & Rajpurkar, P. (2021). CheXtransfer: performance and parameter efficiency of ImageNet models for chest X-Ray interpretation. *Proceedings of the Conference on Health, Inference, and Learning, CHIL '21*, 116–124. https://doi.org/10.1145/3450439.3451867

Kocak, B., Klontzas, M. E., Stanzione, A., Meddeb, A., Demircioğlu, A., Bluethgen, C., Bressem, K. K., Ugga, L., Mercaldo, N., Díaz, O., & Cuocolo, R. (2025). Evaluation metrics in medical imaging AI: fundamentals, pitfalls, misapplications, and recommendations. *European Journal of Radiology Artificial Intelligence*, *3*, 100030. https://doi.org/10.1016/j.ejrai.2025.100030

Kumar, S. (2022). *Covid19-Pneumonia-Normal Chest X-Ray Images [Data set]*. Mendeley Data. https://doi.org/10.17632/dvntn9yhd2.1

Kumar, S., Shastri, S., Mahajan, S., Singh, K., Gupta, S., Rani, R., Mohan, N., & 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*(5), 1464–1480. https://doi.org/10.1002/ima.22770

Liu, F., Tian, Y., Cordeiro, F. R., Belagiannis, V., Reid, I., & Carneiro, G. (2021). Self-supervised Mean Teacher for Semi-supervised Chest X-Ray Classification. In C. Lian, X. Cao, I. Rekik, X. Xu, & P. Yan (Eds.), *Machine Learning in Medical Imaging* (Vol. 12966, pp. 426–436). Springer International Publishing. https://doi.org/10.1007/978-3-030-87589-3_44


TIN: Terapan Informatika Nusantara
Vol 6, No 12, May 2026, page 2345-2360
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i12.9848

Mosquera, C., Ferrer, L., Milone, D. H., Luna, D., & Ferrante, E. (2024). Class imbalance on medical image classification: towards better evaluation practices for discrimination and calibration performance. *European Radiology*, *34*(12), 7895–7903. https://doi.org/10.1007/s00330-024-10834-0

Quiñonez-Baca, L.-C., Ramirez-Alonso, G., Gaxiola, F., Manzo-Martinez, A., Cornejo, R., & Lopez-Flores, D. R. (2025). A Comparative Evaluation of Meta-Learning Models for Few-Shot Chest X-Ray Disease Classification. *Diagnostics*, *15*(18), 2404. https://doi.org/10.3390/diagnostics15182404

Rajaraman, S., Liang, Z., Xue, Z., & Antani, S. (2024). Noise-induced modality-specific pretext learning for pediatric chest X-ray image classification. *Frontiers in Artificial Intelligence*, *7*. https://doi.org/10.3389/frai.2024.1419638

Sajun, A. R., Zualkernan, I., & Sankalpa, D. (2022). Investigating the Performance of FixMatch for COVID-19 Detection in Chest X-rays. *Applied Sciences*, *12*(9), 4694. https://doi.org/10.3390/app12094694

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*, *12*(1), 193–204. https://doi.org/10.1007/s12553-021-00630-x

Sultan, I., Gharaibeh, H., Gharaibeh, A., Lahham, B., Al-Tarawneh, M. K., Al-Qawabah, R., & Nasayreh, A. (2025). LungVisionNet: A Hybrid Deep Learning Model for Chest X-Ray Classification A Case Study at King Hussein Cancer Center (KHCC). *Technologies*. https://api.semanticscholar.org/CorpusID:283000611

Wang, H., Wang, S., Qin, Z., Zhang, Y., Li, R., & Xia, Y. (2021). Triple attention learning for classification of 14 thoracic diseases using chest radiography. *Medical Image Analysis*, *67*, 101846. https://doi.org/10.1016/j.media.2020.101846

Zhang, L., Shi, Z., Chen, M., Chen, Y., Cheng, J., Fan, L., Hong, N., Jia, W., Jiang, G., Ju, S., Li, X., Li, X., Liang, C., Liao, W., Liu, S., Lu, Z., Ma, L., Ren, K., Rong, P., … Jin, Z. (2022). Study design of deep learning based automatic detection of cerebrovascular diseases on medical imaging: a position paper from Chinese Association of Radiologists. *Intelligent Medicine*, *2*(4), 221–229. https://doi.org/10.1016/j.imed.2022.07.001