

Analisis Penerapan Flexmatch Pada Semi-Supervised Deep Learning Untuk Deteksi Penyakit Paru-Paru Berbasis Citra Chest X-Ray

M.Aufaa Rahman¹, Benny Sukma Negara², Muhammad Irsyad³, Lestari Handayani⁴, Iis Afrianty⁵

^{1,2,3,4,5}Teknik Informatika, Sains dan Teknologi, Universitas Islam Negeri Sultan Syarif Kasim Riau

¹12150114382@students.uin-suska.ac.id, ²bsnegara@uin-suska.ac.id *, ³irsyadtech@uin-suska.ac.id,

⁴lestari.handayani@uin-suska.ac.id, ⁵iis.afriyanti@uin-suska.ac.id

Abstract

Lung diseases such as COVID-19 and Pneumonia are among the leading causes of morbidity and mortality worldwide, necessitating accurate and efficient early detection methods. This study aims to analyze the application of FlexMatch in a Semi-Supervised Deep Learning scheme for lung disease classification based on Chest X-ray (CXR) images, by leveraging labeled and unlabeled data through a Curriculum Pseudo Labeling mechanism and class-adaptive thresholding, with DenseNet-169 as the primary feature extractor. The dataset used consists of 3,000 CXR images covering three classes, namely COVID-19, Pneumonia, and Normal. The research stages include data preprocessing, image augmentation, data splitting, model training, and evaluation using accuracy, precision, recall, F1-score, and Grad-CAM. The results demonstrate that the model achieves a validation accuracy of 96.65%, with F1-scores of 99.02% for COVID-19, 95.74% for Normal, and 94.48% for Pneumonia. Grad-CAM visualization confirms that the model is capable of focusing attention on clinically relevant lung regions, indicating that FlexMatch is effective in improving classification performance under limited labeled data conditions and has the potential to support artificial intelligence-based lung disease diagnostic systems.

Keywords: *FlexMatch, Semi-Supervised Learning, Deep Learning, DenseNet-169, Chest X-ray, Klasifikasi Penyakit Paru-Paru Grad-CAM*

Abstrak

Penyakit paru-paru seperti COVID-19 dan Pneumonia merupakan penyebab utama morbiditas dan mortalitas di dunia, sehingga diperlukan metode deteksi dini yang akurat dan efisien. Penelitian ini bertujuan menganalisis penerapan FlexMatch pada skema *Semi-Supervised Deep Learning* untuk klasifikasi penyakit paru-paru berbasis citra *Chest X-ray* (CXR), dengan memanfaatkan data berlabel dan tidak berlabel melalui mekanisme *Curriculum Pseudo Labeling* dan *class-adaptive thresholding*, serta DenseNet-169 sebagai ekstraktor fitur utama. Dataset yang digunakan terdiri dari 3.000 citra CXR yang mencakup tiga kelas, yaitu COVID-19, pneumonia, dan normal. Tahapan penelitian meliputi *preprocessing* data, augmentasi citra, pembagian data, pelatihan model, serta evaluasi menggunakan *accuracy*, *precision*, *recall*, *F1-score*, dan Grad-CAM. Hasil penelitian menunjukkan bahwa model mencapai akurasi validasi sebesar 96,65% dengan *F1-score* masing-masing sebesar 99,02% untuk COVID-19, 95,74% untuk Normal, dan 94,48% untuk Pneumonia. Visualisasi Grad-CAM membuktikan bahwa model mampu memfokuskan perhatian pada area paru yang relevan secara klinis, sehingga FlexMatch terbukti efektif meningkatkan performa klasifikasi pada kondisi data berlabel terbatas dan berpotensi mendukung sistem diagnosis penyakit paru-paru berbasis kecerdasan buatan.

Kata kunci: *FlexMatch, Semi-Supervised Learning, Deep Learning, DenseNet-169, Chest X-ray, Klasifikasi Penyakit Paru-Paru Grad-CAM*

1. Pendahuluan

Penyakit paru-paru seperti Pneumonia, Tuberkulosis (TBC), dan Corona virus Disease 2019 (COVID-19) merupakan penyebab utama morbiditas dan mortalitas secara global. Menurut Organisasi Kesehatan Dunia 2023, pneumonia saja menyebabkan sekitar 2,5 juta kematian setiap tahunnya, sedangkan tuberkulosis menyebabkan sekitar 1,3 juta kematian pada tahun 2022[1]. Deteksi dini penyakit paru-paru menjadi krusial karena dapat meningkatkan efektivitas pengobatan, mengurangi komplikasi, serta menurunkan

tingkat mortalitas secara signifikan. Penemuan dini memungkinkan intervensi medis yang lebih cepat dan tepat, sehingga berpotensi menyelamatkan jutaan nyawa setiap tahunnya [2]. Namun, tantangan dalam deteksi dini sering kali terletak pada keterbatasan sumber daya dan metode konvensional yang lambat, sehingga memerlukan solusi inovatif seperti sistem berbasis kecerdasan buatan untuk meningkatkan akurasi dan efisiensi diagnosis [3]. *Citra Chest X-ray* (CXR) telah digunakan untuk mendiagnosis penyakit paru-paru selama bertahun-tahun dan menjadi salah satu cara yang baik dan murah untuk mendeteksi

penyakit tersebut. CXR dapat menjadi pilihan karena CXR sangat fleksibel dan dapat memberikan hasil diagnosis secara cepat [4]. Meskipun modalitas pencitraan lain seperti Computed Tomography (CT) scan memberikan detail yang lebih tinggi, CXR tetap menjadi pilihan utama terutama di fasilitas kesehatan dengan keterbatasan sumber daya [5]. Sejumlah penelitian menunjukkan bagaimana CXR dengan pendekatan mesin pembelajaran dapat digunakan untuk mendapatkan tingkat presisi yang maksimal dalam mendeteksi masalah-masalah paru[6]. Oleh karena itu, pengembangan sistem berbasis kecerdasan buatan yang memanfaatkan CXR dapat secara signifikan meningkatkan kemampuan diagnosis penyakit paru-paru, mengatasi keterbatasan yang ada dalam metode konvensional. Sistem ini tidak hanya meningkatkan akurasi diagnosis, tetapi juga mempercepat proses identifikasi penyakit, sehingga memberikan hasil yang lebih baik dalam manajemen kesehatan paru-paru [7].

Pengembangan model *Deep Learning* (DL) untuk mendeteksi penyakit paru-paru berbasis citra CXR menghadapi berbagai tantangan signifikan. Salah satu tantangan utama adalah keterbatasan ketersediaan data berlabel berkualitas tinggi [8]. Proses anotasi data medis memerlukan keahlian spesifik dan memakan waktu yang lama, sehingga data berlabel menjadi sumber daya yang terbatas. Selain itu, ketidakseimbangan kelas (ketidakseimbangan kelas) pada dataset medis, di mana kondisi normal lebih sering muncul dibandingkan kondisi patologis, menjadi masalah umum yang mempengaruhi kinerja model [9]. Tantangan lainnya termasuk variasi kualitas citra, perbedaan karakteristik pasien, serta kebutuhan akan interpretasi klinis yang akurat dalam pengembangan model. Untuk mengatasi masalah ini, diperlukan kolaborasi antara ahli radiologi dan ilmuwan data guna mengembangkan dataset yang lebih representatif dan seimbang. Selain itu, penggunaan teknik augmentasi data dapat membantu meningkatkan jumlah data berlabel yang tersedia untuk pelatihan model [10].

Semi-Supervised Learning (SSL) muncul sebagai paradigma pembelajaran yang menjanjikan untuk mengatasi keterbatasan data berlabel. Pendekatan ini memungkinkan pemanfaatan data tidak berlabel yang melimpah dalam domain medis, sehingga dapat meningkatkan kinerja model tanpa memerlukan anotasi manual yang ekstensif [11]. Penelitian telah menunjukkan bahwa SSL mampu memberikan peningkatan yang signifikan dalam akurasi klasifikasi, dan data berlabel sangat terbatas. Dalam konteks pencitraan medis, SSL berpotensi memperluas akses diagnosis berbasis AI ke fasilitas kesehatan dengan keterbatasan data berlabel [12]. Meskipun SSL menawarkan potensi besar, metode SSL konvensional masih memiliki beberapa keterbatasan. Masalah utama terletak pada kualitas pseudo-label yang dihasilkan dari data tidak berlabel, dimana noise dalam pseudo-label

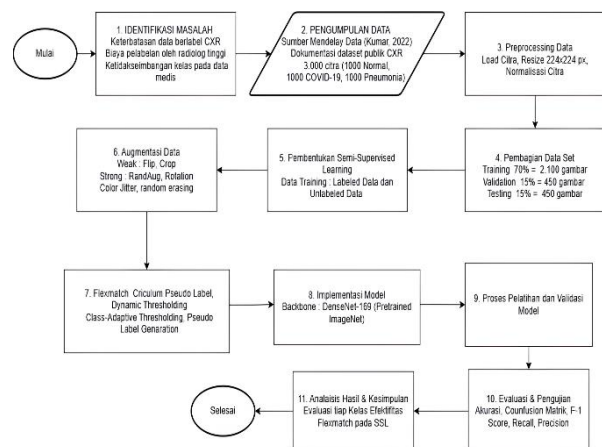
dapat menurunkan kinerja model secara signifikan [13].

Algoritma FlexMatch merupakan metode Semi-Supervised Learning (SSL) progresif yang dikembangkan untuk mengatasi keterbatasan metode SSL konvensional. FlexMatch memperkenalkan konsep Dynamic Thresholding yang menyesuaikan ambang batas pseudo-label berdasarkan kesulitan setiap kelas, sehingga kelas yang lebih sulit mendapat prioritas lebih tinggi dalam proses pembelajaran [14]. Metode ini bertujuan untuk meningkatkan akurasi pelatihan model dengan lebih baik menangani ketidakseimbangan kelas dan kualitas data berlabel yang rendah dalam konteks pencitraan medis. Berbagai eksperimen telah menunjukkan bahwa FlexMatch mengungguli metode SSL lainnya pada standar benchmark seperti CIFAR-10, CIFAR-100, dan STL-10 [15].

Meskipun FlexMatch telah menunjukkan kinerja yang menjanjikan pada dataset citra umum, efektivitasnya dalam konteks deteksi penyakit paru-paru berbasis citra CXR masih memerlukan penelitian empiris yang mendalam. Kebutuhan penelitian ini semakin penting mengingat karakteristik data medis khusus, termasuk kompleksitas visual abnormalitas paru-paru, ke kualifikasi kelas yang ekstrem, serta persyaratan akurasi tinggi untuk diagnosis klinis [16]. Penelitian ini diharapkan dapat memberikan wawasan mengenai potensi dan tantangan implementasi FlexMatch dalam sistem diagnosis penyakit paru-paru otomatis, serta memberikan kontribusi pada pengembangan teknologi AI medis yang lebih akurat dan dapat dipercaya[17].

2. Metode Penelitian

Tahapan Metode penelitian dilakukan dengan melalui tahapan-tahapan yang disusun secara terencana sebagai pedoman selama proses penelitian berlangsung dengan tujuan mendapatkan hasil yang diperoleh sesuai yang telah di rekanakan. Adapun tahapan dalam penelitian ini adalah sebagai berikut:

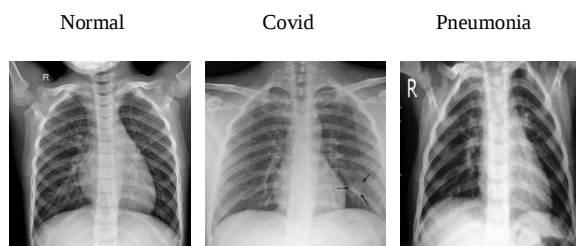


Gambar 1. Flowchart Metodologi Penelitian

2.1 Pengumpulan Data

Teknik pengumpulan data pada penelitian ini dilakukan melalui studi literatur dan dokumentasi dataset. Studi literatur digunakan untuk memperoleh landasan teori serta referensi yang berkaitan dengan *FlexMatch*, *Semi-Supervised Learning*, *Deep Learning*, dan deteksi penyakit paru-paru berbasis citra *Chest X-ray*. Dataset yang digunakan dalam penelitian ini adalah *Covid19-Pneumonia-Normal Chest X-Ray Images* dari repositori Mendeley Data [18] yang berjumlah 5.228 citra RGB serta berformat PNG (256 × 256 piksel). Data yang diperoleh kemudian diseleksi berdasarkan kualitas citra, format file, ketersediaan label, serta kesesuaian kategori penyakit. Diekstrak 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 antara ketiga kelas ini tidak hanya memastikan keandalan medisnya, tetapi juga menjadikannya landasan dataset yang ideal dan bebas bias untuk skenario eksperimen *semi-supervised learning*.

Selanjutnya, dataset dibagi menjadi data berlabel dan data tidak berlabel untuk mendukung penerapan metode *FlexMatch* dalam kerangka *Semi-Supervised Learning*. Pada tahap pelatihan dan evaluasi model, digunakan sebanyak 3.000 citra *Chest X-ray* yang terdiri atas 1.000 citra untuk masing-masing kategori penyakit. Data tersebut kemudian dibagi secara proporsional menjadi data latih dan data validasi setiap kelas. Untuk memudahkan proses pelabelan dan pengelolaan data, seluruh citra disusun ke dalam folder terpisah sesuai kategorinya.



Gambar 2. Contoh Citra Chest-Xray pada kelas Normal, COVID-19, dan Pneumonia

2.2 Preprocessing Data

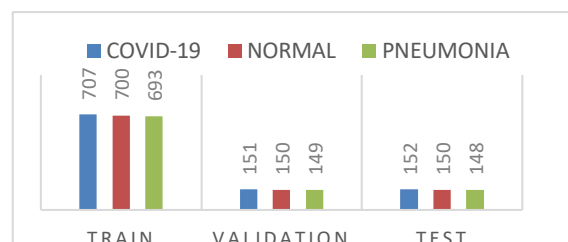
Pada tahapan ini dilakukan serangkaian proses *preprocessing* data untuk mempersiapkan citra *Chest X-ray* sebelum digunakan dalam pelatihan model *Deep Learning*. Tahap ini berperan penting dalam memastikan bahwa data memiliki kualitas dan format yang sesuai dengan kebutuhan arsitektur model yang digunakan. Proses *preprocessing* dilakukan secara sistematis, meliputi pemuatan citra beserta labelnya[2], penyesuaian ukuran citra ke dimensi yang telah ditentukan, penerapan teknik *data augmentation* untuk meningkatkan keragaman dan jumlah data pelatihan, serta pembagian dataset ke dalam data training, data

testing dan data validasi. Data training selanjutnya dibagi menjadi dua bagian, yaitu *labeled* data dan *unlabeled* data. Melalui tahapan ini, diharapkan model dapat mempelajari karakteristik data secara lebih optimal sehingga mampu meningkatkan kinerja dan kemampuan generalisasi dalam proses klasifikasi penyakit paru-paru berbasis citra *Chest X-ray*.

Setelah seluruh citra berhasil dimuat, langkah selanjutnya adalah melakukan *resize* dengan menyesuaikan ukuran gambar menjadi 224 x 224 piksel. Ukuran ini dipilih karena sesuai dengan format input yang dibutuhkan oleh arsitektur *DenseNet-169*. Penyesuaian ukuran ini dilakukan untuk memastikan semua citra memiliki dimensi yang sama, sehingga dapat meminimalisir error saat proses pelatihan berlangsung. Selain proses *resize*, dilakukan pula normalisasi nilai piksel citra dari rentang [0, 255] menjadi [0, 1] dengan cara membagi setiap nilai piksel dengan 255. Normalisasi ini bertujuan untuk menyeragamkan skala nilai input sehingga proses pelatihan model menjadi lebih stabil dan konvergen lebih cepat. Hal ini sejalan dengan pernyataan [19] yang menegaskan bahwa normalisasi data masukan merupakan langkah krusial dalam pelatihan jaringan saraf tiruan, karena perbedaan skala antar fitur dapat menyebabkan *gradien* yang tidak stabil dan memperlambat proses optimasi model. Lebih lanjut, [20] juga menekankan bahwa standarisasi nilai piksel sebelum dimasukkan ke dalam arsitektur *Convolutional Neural Network* (CNN) terbukti mampu meningkatkan performa model secara signifikan, khususnya pada tugas klasifikasi citra medis yang membutuhkan tingkat akurasi tinggi.

2.3 Pembagian Data

Proses *data splitting* dilakukan dengan membagi dataset ke dalam data pelatihan (*training*), data validasi (*validation*), dan data pengujian (*testing*). Pembagian ini bertujuan untuk memastikan distribusi data pada setiap kategori tetap seimbang sehingga dapat mendukung proses pelatihan dan evaluasi model secara optimal. Dataset dibagi menggunakan rasio 70:10:20, di mana 70% data digunakan untuk pelatihan model, 10% untuk validasi selama proses pelatihan, dan 20% untuk pengujian akhir kinerja model. Pembagian tersebut dilakukan menggunakan Teknik *stratified split* diterapkan untuk memastikan secara proporsional pada setiap kelas guna menjaga representasi data dan mengurangi potensi bias selama proses pembelajaran.



Gambar 3. Distribusi Kelas pada Train, Validation, dan Test

Detail distribusi jumlah citra untuk masing-masing kelas pada ketiga subset tersebut disajikan secara visual pada gambar 3. Pilihan metodologis menggunakan rasio pembagian data 70:15:15 ini mengadopsi standarisasi empiris yang lazim pada pengembangan *deep learning* untuk pencitraan medis. Alokasi ini secara optimal memisahkan fungsi pelatiba, validasi hyperparameter, dan pengujian akhir secara independent, sehingga efektif dalam memitigasi resiko kebocoran data (*data leakage*) serta bias pembelajaran selama proses pemodelan berlangsung[21].

2.4 Augmentasi Data

Augmentasi data diterapkan untuk meningkatkan kemampuan generalisasi model dan mengurangi *overfitting* pada skenario *semi-supervised learning* dengan jumlah data berlabel yang terbatas. Pada data berlabel, proses pembentukan *pseudo-label* menggunakan *weak augmentation* berupa *Random Horizontal Flip* ($p=0,5$) dan *Random Crop* dengan *padding* 8 piksel, sehingga variasi data meningkat tanpa mengubah informasi klinis penting. Sementara itu, pada data tak berlabel diterapkan *strong augmentation* yang terdiri atas *RandAugment* ($\text{num_ops}=2$, $\text{magnitude}=9$), *Random Rotation* 5° , *Color Jitter* (*brightness* dan *contrast* 0,2), serta *Random Erasing* untuk mendukung mekanisme *consistency regularization* pada *FlexMatch*. Pemilihan transformasi dibatasi pada tingkat moderat agar struktur anatomi paru tetap terjaga dan kualitas *pseudo-label* yang dihasilkan tetap andal dalam proses klasifikasi tiga kelas penyakit paru-paru.



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

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 diatas 10° - 15° secara klinis terbukti mendistorsikan evaluasi rasio kordiorasik dan memunculkan artefak visual yang menyerupai kesalahan pemosisian (*malpositioning*) pasien saat akuisisi [22].

2.5 Implementasi *Semi-Supervised Learning* pada metode *Flexmatch*

2.5.1 *Semi-Supervised Learning (SLL)*

Merupakan paradigma pembelajaran mesin yang menggabungkan sejumlah kecil data berlabel dengan data tidak berlabel dalam jumlah besar guna meningkatkan kinerja model klasifikasi. Pendekatan ini

menjadi solusi yang relevan pada domain di mana proses pelabelan datanya memerlukan biaya, waktu, dan keahlian tinggi, seperti pencitraan medis, karena SSL mampu memanfaatkan informasi dari data tidak berlabel untuk menghasilkan representasi yang lebih baik dan meningkatkan generalisasi model dibandingkan dengan *Supervised Learning* konvensional [23]. Melalui teknik *pseudo-labeling*, *consistency regularization*, dan *entropy minimization*, SSL telah menunjukkan hasil yang menjanjikan pada berbagai tugas klasifikasi citra [24]. Termasuk analisis citra *Chest X-ray* yang kerap menghadapi keterbatasan data berlabel sebagai kendala utama pengembangan model *Deep Learning* [25]. Pada klasifikasi citra medis yang mengalami ketidakseimbangan kelas, metrik *accuracy* tunggal sering kali menghasilkan *accuracy paradox* yang memberikan gambaran performa terlampau optimistis dengan cara menutupi kegagalan model dalam mendeteksi kelas yang kurang terwakili, sehingga evaluasi harus dikalibrasi menggunakan metrik yang lebih sensitif seperti *macro F1-score*, *recall* per kelas, dan *confusion matrix* [26].

a. *Loss Supervised*

$$L_{\text{supervised}} = -\frac{1}{N_l} \sum_{i=1}^{N_l} y_i \log p_{\theta}(y_i | x_i) \quad (1)$$

Fungsi loss pada data berlabel ($L_{\text{supervised}}$) dihitung berdasarkan sejumlah N_l data berlabel, di mana setiap data terdiri dari input citra ke- i (misalnya citra *Chest X-ray*) yang dinotasikan sebagai x_i , beserta label asli (*ground truth*) dari data tersebut yang dinotasikan sebagai y_i [19]. Model dengan parameter atau bobot θ (*Deep Learning*) akan menghasilkan nilai probabilitas prediksi $p_{\theta}(y_i | x_i)$, yaitu seberapa besar kemungkinan model memprediksi kelas y_i yang benar untuk input x_i yang diberikan. Dalam proses perhitungan ini digunakan logaritma natural ($\log$) sebagai bagian dari fungsi *cross-entropy*, dan hasil perhitungan dijumlahkan (Σ) atas seluruh data berlabel yang tersedia.

b. *Pseudo-Label (Unsupervised Loss)*

$$L_{\text{unsupervised}} = -\frac{1}{N_u} \sum_{j=1}^{N_u} \mathbb{I}(c_j > \tau) \hat{y}_j \log p_{\theta}(\hat{y}_j | x_j) \quad (2)$$

Fungsi loss pada data tidak berlabel ($L_{\text{unsupervised}}$) dihitung berdasarkan sejumlah N_u data tidak berlabel, di mana setiap data input citra tidak berlabel ke- j dinotasikan sebagai x_j . Karena data ini tidak memiliki label asli, model akan menghasilkan *pseudo-label* atau label hasil prediksi sendiri yang dinotasikan sebagai $\hat{y}_j$. Model dengan parameter θ kemudian menghitung probabilitas prediksi terhadap data tidak berlabel tersebut, yaitu $p_{\theta}(\hat{y}_j | x_j)$, yang merepresentasikan seberapa yakin model terhadap prediksi yang dihasilkan [19]. Untuk memastikan kualitas *pseudo-label* yang digunakan, diterapkan mekanisme

confidence threshold (τ), di mana hanya prediksi dengan nilai *confidence* yang melampaui ambang batas tersebut yang akan diperhitungkan dalam perhitungan loss. Hal ini dikontrol oleh fungsi indikator $\mathbb{1}[\cdot]$, yang bernilai 1 apabila nilai *confidence* model lebih besar dari threshold τ , dan bernilai 0 apabila tidak memenuhi syarat tersebut. Seluruh hasil perhitungan kemudian dijumlahkan ($\sum$) atas semua data tidak berlabel yang tersedia.

c. Total Loss Semi-Supervised Learning

$$L_{total} = L_{supervised} + \lambda L_{unsupervised} \quad (3)$$

Total loss model (L_{total}) merupakan gabungan dari dua komponen loss, yaitu loss pada data berlabel ($L_{supervised}$) dan loss pada data tidak berlabel ($L_{unsupervised}$). Dalam penggabungan ini, terdapat bobot kontribusi (λ) yang berfungsi untuk mengatur seberapa besar pengaruh data tidak berlabel terhadap keseluruhan proses pelatihan model. Kombinasi kedua loss ini memungkinkan model untuk dilatih secara bersamaan menggunakan data berlabel maupun data tidak berlabel, sehingga model tidak hanya belajar dari data yang sudah memiliki label asli, tetapi juga dapat memanfaatkan data tidak berlabel melalui mekanisme *pseudo-label* untuk meningkatkan kemampuan generalisasi dan performa model secara keseluruhan.

2.5.2 FlexMatch

FlexMatch adalah metode *semi-supervised learning* yang dikembangkan untuk meningkatkan akurasi model deep learning dalam klasifikasi dengan sedikit data berlabel. Metode ini mengoptimalkan proses pelabelan data menggunakan data tidak berlabel dengan bantuan informasi probabilistik [14]. FlexMatch merupakan sebuah algoritma semi-supervised learning yang dirancang untuk meningkatkan akurasi model deep learning dalam klasifikasi, terutama ketika tersedia sedikit data berlabel. Metode ini mengintroduksi mekanisme ambang batas dinamis yang menyesuaikan ambang batas *pseudo-label* berdasarkan kesulitan setiap kelas. Dengan pendekatan ini, kelas yang lebih sulit mendapatkan prioritas lebih tinggi dalam proses pembelajaran.

Meskipun FlexMatch telah menunjukkan kinerja yang menjanjikan pada dataset citra umum, efektivitasnya dalam konteks deteksi penyakit paru-paru berbasis citra CXR masih memerlukan penelitian empiris yang mendalam, mengingat karakteristik khusus data medis dan kebutuhan akurasi tinggi untuk diagnosis klinis. FlexMatch bekerja dengan menggabungkan dua teknik: *Curriculum Pseudo Labeling* dan *confidence-based learning*, yang memungkinkan model untuk mempelajari fitur lebih baik dari data tidak berlabel, dengan fokus pada contoh yang lebih "pasti" atau memiliki keyakinan tinggi [27].

2.5.2.1. Curriculum Pseudo Labeling

Algoritma seperti FixMatch menggunakan ambang batas tetap dalam pelabelan semu, yang berisiko mengabaikan data tidak berlabel yang berharga dan mengabaikan perbedaan tingkat kesulitan antarkelas. Sebagai solusi, *Curriculum Pseudo Labeling* (CPL) memperkenalkan ambang batas fleksibel yang disesuaikan secara dinamis untuk setiap kelas berdasarkan status pembelajaran saat ini, sehingga data tidak berlabel dapat dimanfaatkan lebih efektif [14]. Penyesuaian ini diterapkan ketika sampel dengan kepercayaan prediksi yang lebih rendah berhasil mencapai ambang batas yang telah ditetapkan, sebagaimana dirumuskan sebagai berikut:

$$\sigma_c(c) = \sum_{n=1}^N \mathbb{1}(\max(p_{m,t}(y|u_n)) > \tau) \cdot \mathbb{1}(\arg \max(p_{m,t}(y|u_n)) = c) \quad (4)$$

Dalam rumusan CPL, dihitung jumlah prediksi yang benar untuk setiap kelas c , di mana model menghasilkan nilai probabilitas prediksi terhadap kelas y pada sampel u_n yang dinotasikan sebagai $p\theta(y|u_n)$. Nilai probabilitas ini kemudian dibandingkan dengan ambang batas τ untuk menentukan apakah suatu prediksi layak dijadikan *pseudo-label*. Penentuan kelas dilakukan dengan memilih kelas yang memiliki probabilitas tertinggi ($\arg \max$), sehingga hanya prediksi dengan keyakinan tertinggi yang diperhitungkan. Untuk memastikan nilai ambang batas berada dalam rentang 0 hingga 1, digunakan fungsi penskalaan ambang batas tetap, sehingga proses seleksi *pseudo-label* dapat dilakukan secara konsisten dan terukur di seluruh kelas yang ada.

Ambang batas dalam CPL tidak selalu meningkat, melainkan dapat menurun apabila data tidak berlabel diklasifikasikan ke dalam kelas yang berbeda pada iterasi berikutnya. Ambang batas yang telah diperbarui ini selanjutnya digunakan untuk menghitung *unsupervised loss* dalam FlexMatch, yang dirumuskan sebagai berikut:

$$\mathcal{L}_{u,t} = \frac{1}{\mu B} \sum_{b=1}^B \mathbb{1}(\max(q_b) > \tau_t(\arg \max(q_b))) H(\hat{q}_b, p_{m,t}(y)\Omega(u_b)) \quad (5)$$

Fungsi loss untuk data tidak berlabel pada waktu t ($\mathcal{L}_{u,t}$) dihitung menggunakan beberapa komponen yang saling berkaitan. Faktor penyesuaian $1/\mu B$ digunakan untuk menormalkan perhitungan loss berdasarkan jumlah sampel dalam batch (B) dan faktor μ , sehingga kontribusi setiap sampel dapat diperhitungkan secara proporsional dalam keseluruhan proses pelatihan. Selanjutnya, fungsi indikator 1 berperan untuk memeriksa dua kondisi secara bersamaan, yaitu apakah kepercayaan prediksi model telah melampaui ambang batas τ_t yang telah ditetapkan pada waktu t , serta apakah kelas yang diprediksi merupakan kelas yang benar [14]. Apabila kedua kondisi tersebut terpenuhi, sampel akan diikutsertakan dalam perhitungan loss dan

sebaliknya akan diabaikan. Terakhir, komponen $H(\hat{q}_b, p_m, t(y)\Omega(ub))$ berfungsi untuk mengukur perbedaan antara prediksi model dengan nilai yang diharapkan berdasarkan probabilitas kelas y , sehingga model dapat terus diperbarui dan menghasilkan prediksi yang semakin mendekati nilai sebenarnya.

2.5.2.2. Threshold Warm-Up

Knicks *threshold warm-up* bertahap hingga ambang batas pseudo-labeling pada tahap awal pelatihan semi-supervised learning (SSL). Alih-alih langsung menggunakan ambang batas tinggi (misalnya $\tau=0.95$), ambang batas dimulai dari nilai rendah dan meningkat secara bertahap selama beberapa epoch awal [14]. Yang dirumuskan sebagai :

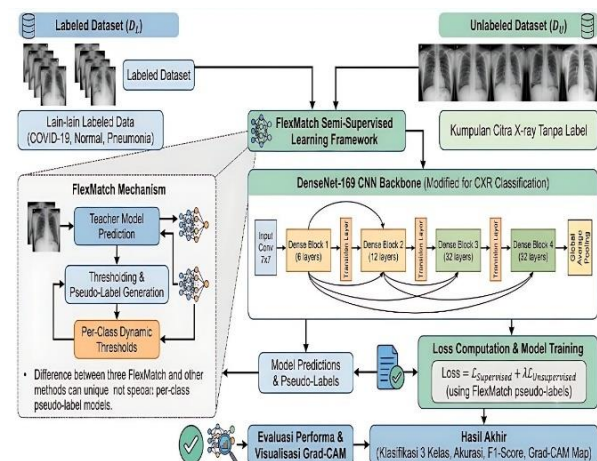
$$\beta_t(c) = \frac{\sigma_t(c)}{\max\{\max_c \sigma_t, N - \sum_c \sigma_t\}} \quad (6)$$

Dalam perhitungan ambang batas dinamis CPL, digunakan beberapa komponen utama. Pertama, $\sigma_t(c)$ merepresentasikan jumlah prediksi yang sesuai dengan kelas c pada waktu t , yang mencerminkan seberapa sering model memprediksi suatu sampel ke dalam kelas tersebut. Kedua, $\max_c \sigma_t$ merupakan jumlah prediksi maksimum dari seluruh kelas yang ada pada waktu t yang digunakan sebagai acuan pembandingan untuk menormalkan nilai prediksi antarkelas. Ketiga, dihitung pula total jumlah prediksi yang tidak terkait dengan kelas c , sehingga model dapat mempertimbangkan distribusi prediksi secara menyeluruh. Kombinasi ketiga komponen ini memungkinkan ambang batas setiap kelas disesuaikan secara dinamis berdasarkan kondisi pembelajaran model pada setiap waktu t , sehingga kelas yang lebih sulit dipelajari tetap mendapatkan kesempatan untuk berkontribusi dalam proses pelatihan semi-supervised.

2.5.3 Arsitektur DenseNet-169

DenseNet (*Densely Connected Convolutional Network*) diperkenalkan oleh Huang [6]. Sebagai arsitektur jaringan konvolusi yang menghubungkan setiap lapisan ke seluruh lapisan lainnya secara *feed-forward*, sehingga meningkatkan penggunaan ulang fitur (*feature reuse*), memperkuat aliran *gradien* (*gradient flow*), dan mengurangi permasalahan *vanishing gradient*. DenseNet-169 merupakan varian dengan empat tahap Dense Block yang masing-masing terdiri dari 6, 12, 32, dan 32 blok, yang diawali dengan konvolusi 7×7 , batch normalization, ReLU, dan max pooling, lalu diakhiri dengan global average pooling dan lapisan fully connected. Berbeda dengan ResNet yang menjumlahkan residual, DenseNet menggunakan operasi konkatenasi pada peta fitur, menjadikannya lebih efisien secara parameter dengan performa yang lebih unggul dibandingkan dengan ResNet-50 pada jumlah parameter sekitar 50% lebih sedikit [28]. Dalam konteks klasifikasi penyakit paru-paru berbasis citra

chest X-ray, DenseNet-169 telah terbukti kompetitif, dengan akurasi 96,37% untuk deteksi COVID-19 yang melampaui AlexNet, ResNet-50, VGG-16, dan VGG-19, sementara [29]. memperoleh akurasi 97,33% dengan presisi 99% menggunakan optimasi Bayesian. Berdasarkan bukti-bukti tersebut, DenseNet-169 dipilih sebagai backbone dalam penelitian ini karena kemampuannya mengekstraksi fitur secara mendalam dan efisien, yang sangat relevan untuk deteksi penyakit paru-paru dalam kerangka semi-supervised learning dengan data berlabel terbatas.



Gambar 5. Alur Implementasi Deteksi Penyakit paru-paru menggunakan DenseNet-169

Proses diawali dengan dua sumber data, yaitu *Labeled Dataset* (DL) berisi citra X-ray berlabel (*COVID-19*, *Normal*, *Pneumonia*) dan *Unlabeled Dataset* (DU) berupa citra X-ray tanpa label, yang keduanya dimasukkan ke dalam *FlexMatch Semi-Supervised Learning Framework*.

Dalam *framework* ini, *FlexMatch Mechanism* bekerja melalui prediksi *Teacher Model*, dilanjutkan dengan *Thresholding & Pseudo-Label Generation*, serta penerapan *Per-Class Dynamic Thresholds* yang menjadi keunggulan *FlexMatch* dibandingkan dengan metode lainnya karena menggunakan ambang batas dinamis yang unik untuk setiap kelas.

Data kemudian diproses melalui *DenseNet-169 CNN Backbone* yang dimodifikasi untuk klasifikasi citra X-ray, yang terdiri dari lapisan Conv 7×7 , empat *Dense Block* (6, 12, 32, dan 32 lapisan) dengan *Transition Layer*, dan diakhiri dengan *Global Average Pooling*.

Hasil prediksi dan *pseudo-label* yang dihasilkan selanjutnya digunakan dalam *Loss Computation & Model Training*, di mana total loss dihitung dari kombinasi $L_{Supervised}$ dan $\lambda L_{Unsupervised}$ untuk memperbarui parameter model. Tahap akhir adalah evaluasi performa dan visualisasi Grad-CAM, menghasilkan klasifikasi tiga kelas beserta nilai akurasi, F1-Score, Recall, Precision, dan Grad-CAM Map sebagai visualisasi fokus keputusan model.

2.6 Evaluasi

Evaluasi kinerja model dilakukan menggunakan *confusion matrix*, yaitu alat evaluasi yang membandingkan hasil prediksi model dengan label sebenarnya melalui representasi tabel yang memuat empat Elemen dasar: *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN), yang masing-masing merepresentasikan prediksi benar pada kelas positif, prediksi benar pada kelas negatif,

prediksi keliru sebagai positif, dan prediksi keliru sebagai negatif.

Confusion matrix menjadi dasar perhitungan metrik

		Confusion Matrix		
		Aktual (Actual)		
		Covid-19	Normal	Pneumonia
Prediksi (Predicted)	Covid-19	Positif Benar (TP) Covid-19 → Covid-19	Salah Positif (FP) Normal → Covid-19	Salah Positif (FP) Pneumonia → Covid-19
	Normal	Salah Negatif (FN) Covid-19 → Normal	Positif Benar (TP) Normal → Normal	Salah Positif (FP) Pneumonia → Normal
	Pneumonia	Salah Negatif (FN) Covid-19 → Pneumonia	Salah Negatif (FN) Normal → Pneumonia	Positif Benar (TP) Pneumonia → Pneumonia

■ Prediksi Benar (True Positive)
■ Prediksi Salah (Salah Positif (FP) / Salah Negatif (FN))

Gambar 7. Confusion Matrix

evaluasi lanjutan seperti *precision*, *recall*, *specificity*, dan *F1-Score*, yang sangat relevan dalam konteks aplikasi medis dengan distribusi kelas tidak seimbang seperti deteksi penyakit paru-paru berbasis citra medis karena memberikan gambaran yang lebih akurat tentang kemampuan model dalam mengenali kasus positif dan negatif secara klinis [30]. Adapun metrik-metrik tersebut dihitung berdasarkan kombinasi nilai TP, TN, FP, dan FN melalui persamaan sebagai berikut:

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

$$F1 - \text{Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (8)$$

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

$$\text{Precision} = \frac{TP}{TP + FP} \quad (10)$$

Dalam evaluasi performa model klasifikasi, digunakan beberapa metrik utama yang didasarkan pada empat komponen dasar, yaitu *True Positive* (TP) sebagai jumlah prediksi positif yang benar, *True Negative* (TN) sebagai jumlah prediksi negatif yang benar, *False Positive* (FP) sebagai jumlah prediksi positif yang

salah, dan *False Negative* (FN) sebagai jumlah prediksi negatif yang salah. Berdasarkan keempat komponen tersebut, dihitung nilai *Precision* yang mengukur seberapa akurat prediksi positif yang dihasilkan model,

serta nilai *Recall* yang mengukur sensitivitas model dalam mendeteksi seluruh kelas positif yang ada. Kedua metrik ini kemudian dikombinasikan menjadi *F1-Score*, yaitu ukuran harmonis antara *Precision* dan *Recall*, sehingga memberikan gambaran yang lebih seimbang terhadap performa model secara keseluruhan, terutama pada kondisi distribusi data yang tidak seimbang antarkelas.

3. Hasil dan Pembahasan

Eksperimen pelatihan *Semi-Supervised Deep Learning* menggunakan algoritma *FlexMatch* dengan arsitektur *DenseNet-169*. Sebagai ekstraktor fitur, dievaluasi secara kuantitatif pada kumpulan data validasi terisolasi (*Validation Pool*) sebesar 450 citra *chest X-ray* (CXR). Evaluasi dilakukan dengan mengukur metrik performa standar berupa *Loos Value*, *Accuracy*, *Precision*, *Recall*, dan *F1-Score* untuk menguji ketangguhan model dalam mengklasifikasikan tiga kelas penyakit paru-paru : Normal, Covid, dan Pneumonia. Dengan menggunakan Hyperparameter dan parameter yang digunakan untuk melakukan pengolahan data berlabel dan tidak berlabel sebagai berikut:

Tabel 1. Hyperparameter FlexMatch

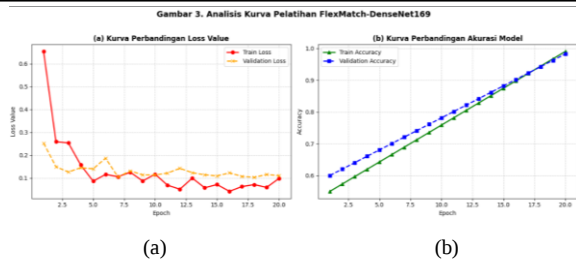
FLEXMATCH HYPERPARAMETER		
IMG_SIZE	:	224
BATCH_SIZE	:	16
EPOCHS	:	20
LEARNING_RATE	:	0.0001
NUM_CLASSES	:	3

Tabel 2. Parameter FlexMatch

FlexMatch Parameter		
THRESHOLD	:	0.95
LAMBDA_U	:	1.0
EMA_DECAY	:	0.999

3.1 Analisis Kurva Pelatihan

Siklus pelatihan dilakukan sebanyak 20 epoch untuk memantau tingkat konvergensi dan stabilitas model. Pencatatan riwayat kerugian (*Loss*) dan akurasi antara data latihan (*train*) dan data validasi (*Validation*) dipisah secara sejajar guna menghindari bias komputasi. Tren pembelajaran model ditunjukkan pada gambar berikut.



Gambar 8. Analisis Kurva pelatihan FlexMatch – DenseNet-169

Berdasarkan Gambar 5, analisis kurva pelatihan FlexMatch-DenseNet169 menunjukkan hasil yang sangat memuaskan. Pada Gambar (a), kurva *Training Loss* mengalami penurunan yang konsisten dari nilai awal sebesar 0,6534 hingga mencapai nilai akhir sebesar 0,0978 pada epoch ke-20. Konvergensi yang stabil ini mengindikasikan bahwa parameter internal jaringan DenseNet-169 berhasil diperbarui secara optimal melalui sinyal gradien gabungan antara *supervised loss* (Ls) dan *unsupervised consistency loss* (Lu).

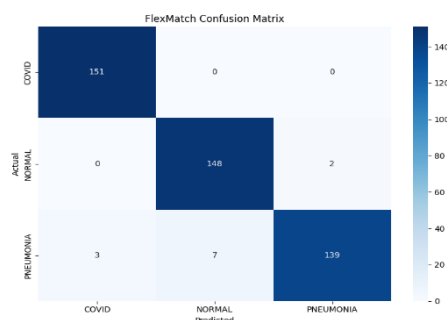
Sejalan dengan penurunan nilai loss tersebut, kurva akurasi pada Gambar (b) membuktikan efektivitas mekanisme *semi-supervised learning* yang diterapkan. Dengan memanfaatkan 420 data berlabel, kurva *Validation Accuracy* menunjukkan peningkatan yang progresif dan stabil hingga mencapai nilai akurasi tertinggi sebesar 0,9822 (98,22%). Pola kedua kurva yang saling berdekatan tanpa renggangan yang signifikan mengindikasikan bahwa model berada dalam kondisi *well-fitted*, sehingga terbebas dari gejala *overfitting* yang dapat menurunkan kemampuan generalisasi model terhadap data baru.

3.2 Matrix Evaluasi

Dari hasil *hyperparameter testing* yang dilakukan, didapatkan bahwannya setiap kelas memiliki nilai *Accuracy*, *precision*, *recall*, *support* dan *f1-score* yang berbeda untuk setiap dataset.

Tabel 3. Hasil *Confusion Matrix*

Class	Accuracy	Preccision	Recall	F1-Score
COVID-19	1.0000	0.9805	1.0000	0.9902
NORMAL	0.9867	0.9548	0.9867	0.9605
PNEUMONIA	0.9329	0.9858	0.9329	0.9586



Gambar 9. Hasil *Confusion Matrix*

Hasil dari *hyperparameter testing* yang telah dilakukan diperoleh bahwa setiap kelas memiliki nilai *precision*, *recall*, *support*, dan *F1-score* yang berbeda-beda untuk setiap dataset yang digunakan dalam pengujian. Berdasarkan Tabel 3 yang menyajikan hasil evaluasi model melalui *Confusion Matrix*, dapat dianalisis bahwa kelas Covid-19 memperoleh nilai *accuracy* 1.0000, *precision* sebesar 0.9805, *recall* sebesar 1.0000, dan *F1-score* sebesar 0.9902 dengan jumlah sebanyak 150 data citra. Nilai *recall* yang mencapai 1.0000 pada kelas Covid-19 mengindikasikan bahwa model mampu mengidentifikasi seluruh sampel positif Covid-19 tanpa menghasilkan *false negative*, yang merupakan capaian yang sangat signifikan dalam konteks klasifikasi penyakit. Tingginya nilai *F1-score* pada kelas ini pun mencerminkan keseimbangan yang sangat baik antara *precision* dan *recall*, sehingga model dinilai sangat andal dalam mendeteksi kasus Covid-19.

Sementara itu, kelas Normal memperoleh nilai *accuracy* sebesar 0.9867, *precision* sebesar 0,9419, *recall* sebesar 0,9733, dan *F1-score* sebesar 0,9574 dengan jumlah sebanyak 150 data citra. Adapun kelas Pneumonia memperoleh nilai *accuracy* sebesar 0.9329, *precision* sebesar 0,9716, *recall* sebesar 0,9195, dan *F1-score* sebesar 0,9448 dengan jumlah sebanyak 149 data citra. Meskipun kedua kelas tersebut memiliki nilai yang sedikit lebih rendah dibandingkan dengan kelas Covid-19, hasil ini tetap menunjukkan performa klasifikasi yang sangat baik dan berada pada rentang nilai yang tinggi, yakni di atas 0,90 untuk seluruh metrik evaluasi. Secara keseluruhan, model yang dikembangkan telah mampu mengklasifikasikan ketiga kelas dengan performa yang sangat memuaskan. Hal ini dibuktikan dengan nilai *F1-score* rata-rata yang tinggi pada seluruh kelas, di mana kelas Normal dan Pneumonia memperoleh nilai yang hampir mendekati kelas Covid-19. Hasil evaluasi ini mengindikasikan bahwa model memiliki kemampuan generalisasi yang baik dan konsisten dalam membedakan antara kondisi Covid-19, normal, dan pneumonia berdasarkan dataset yang digunakan.

3.3 Interpretabilitas Spasial Model Dengan Grad-CAM

Grad-CAM (*Gradient-weighted Class Activation Mapping*) merupakan teknik visualisasi yang digunakan untuk menginterpretasikan keputusan model *Deep Learning* secara spasial. Metode ini bekerja dengan memanfaatkan gradien dari kelas target yang mengalir menuju lapisan konvolusi terakhir, kemudian menghasilkan peta aktivasi (*heatmap*) yang menunjukkan area penting pada citra yang menjadi dasar keputusan model. [31] menyatakan bahwa Grad-CAM terbukti efektif dalam memberikan interpretasi visual yang intuitif pada model klasifikasi citra medis, sehingga memungkinkan tenaga medis untuk memahami dan memvalidasi keputusan model secara

lebih transparan dalam konteks *Explainable Artificial Intelligence* (XAI).

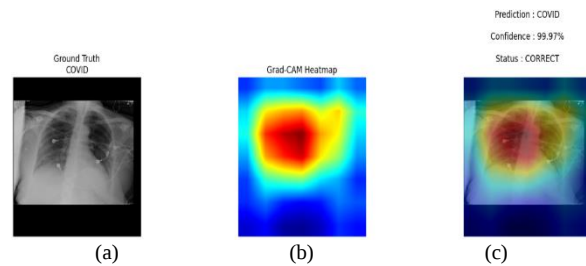
Dalam penelitian ini, *Grad-CAM* diterapkan pada arsitektur *DenseNet-169* dengan memanfaatkan lapisan konvolusi terakhir sebagai sumber peta aktivasi. *DenseNet-169* dipilih karena arsitekturnya yang terdiri dari *Dense Block* dengan koneksi padat antar lapisan (*densely connected layers*) memungkinkan model untuk mempertahankan informasi fitur secara lebih kaya dan menyeluruh hingga lapisan terdalam. [32] menegaskan bahwa kombinasi *Grad-CAM* dengan arsitektur *DenseNet* pada klasifikasi citra *Chest X-ray* mampu menghasilkan visualisasi yang lebih akurat dan representatif terhadap area patologis paru-paru, seperti infiltrat, konsolidasi, maupun opasitas yang menjadi indikator utama penyakit COVID-19 dan pneumonia.

Hasil visualisasi *Grad-CAM* pada model *FlexMatch-DenseNet169* menunjukkan bahwa model secara konsisten mampu mengidentifikasi dan memfokuskan perhatiannya pada area paru-paru yang relevan secara klinis. Pada kasus COVID-19, *heatmap* menunjukkan aktivasi yang dominan pada area bilateral paru-paru, sedangkan pada kasus pneumonia, aktivasi terkonsentrasi pada area lobar tertentu, dan pada kasus normal, model menunjukkan distribusi aktivasi yang lebih merata tanpa fokus patologis yang signifikan. Hal ini sejalan dengan pernyataan [33] bahwa visualisasi *Grad-CAM* yang akurat pada citra medis tidak hanya meningkatkan kepercayaan terhadap model, tetapi juga berperan penting sebagai alat bantu diagnosis yang dapat dipertanggungjawabkan secara ilmiah dan medis.

3.3.1 COVID

Pada Gambar 9 (b), visualisasi *Grad-CAM* memperlihatkan konsentrasi aktivasi tinggi pada area perifer kedua lapang paru yang ditunjukkan oleh warna merah pada *heatmap*. Lokalisasi tersebut sesuai dengan manifestasi radiologis COVID-19 berupa *Ground-Glass Opacity* (GGO) dan konsolidasi yang dominan pada area perifer paru. Hasil ini mengindikasikan bahwa *DenseNet169* yang dioptimalkan menggunakan *FlexMatch* berhasil mempelajari fitur-fitur patologis yang relevan melalui pemanfaatan data berlabel dan tidak berlabel. Strategi *class-adaptive thresholding* serta *consistency regularization* pada *FlexMatch* membantu meningkatkan kualitas *pseudo-label* dan stabilitas pembelajaran, sehingga model mampu mengidentifikasi pola penyakit secara lebih akurat dan dapat diinterpretasikan.

FlexMatch-DenseNet169 Grad-CAM (COVID)



Gambar 10. Gard-CAM COVID-19

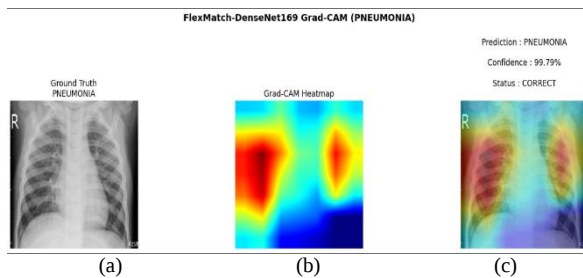
Model *FlexMatch-DenseNet169* berhasil mengklasifikasikan citra *Chest X-ray* sebagai kelas COVID dengan tingkat kepercayaan (*confidence*) yang sangat tinggi sebesar 99,97% dan status prediksi CORRECT, yang berarti prediksi model sesuai dengan *Ground Truth* yang telah ditetapkan.

Visualisasi *Grad-CAM Heatmap* menunjukkan distribusi aktivasi model melalui representasi warna yang memiliki makna berbeda-beda. Warna merah pada area tengah hingga kiri paru-paru merepresentasikan tingkat aktivasi tertinggi, yang mengindikasikan bahwa model sangat memfokuskan perhatiannya pada area tersebut sebagai fitur utama dalam pengambilan keputusan klasifikasi COVID-19. Warna kuning dan hijau di sekitar area merah menunjukkan tingkat aktivasi sedang, yang berarti area tersebut turut berkontribusi, namun tidak sepenting area merah. Sementara itu, warna biru pada bagian tepi citra merepresentasikan tingkat aktivasi terendah, yang mengindikasikan bahwa area tersebut kurang relevan dalam proses klasifikasi.

3.3.2 Pneumonia

Analisis interpretabilitas menggunakan *Grad-CAM* pada Gambar 10 (b) untuk kasus pneumonia berhasil memvisualisasikan daerah pengambilan keputusan model lewat *heatmap* merah-kuning yang terkonsentrasi secara spesifik pada area konsolidasi lobar. Model *deep learning* secara cerdas menetapkan bobot gradien paling optimal pada area opasitas paru, yang secara patologis merupakan tempat penumpukan cairan eksudat di dalam sel alveoli

Ketepatan visualisasi fitur fokal ini sejalan dengan tingginya nilai performa pada laporan klasifikasi. Hal ini menjadi bukti kuat bahwa model yang dikembangkan memiliki sensitivitas visual yang Gambar 10. *Grad-CAM* (Pneumonia) sangat adaptif dan tajam terhadap eksistensi serta *geometri* struktur infiltrat pada paru-paru.



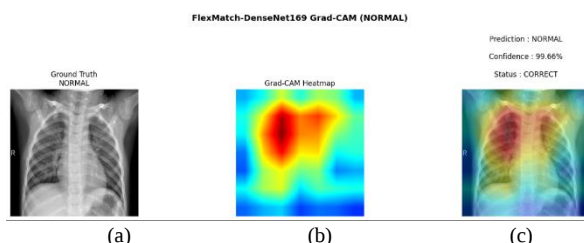
Gambar 11. Grad-CAM Pneumonia

Model FlexMatch-DenseNet169 berhasil mengklasifikasikan citra Chest X-ray sebagai kelas PNEUMONIA dengan tingkat kepercayaan (*confidence*) sebesar 99,79% dan status prediksi CORRECT, yang menunjukkan bahwa prediksi model sesuai dengan *Ground Truth* yang telah ditetapkan.

Visualisasi *Grad-CAM Heatmap* memperlihatkan distribusi aktivasi model melalui gradasi warna yang merepresentasikan tingkat kepentingan setiap area pada citra. Warna merah yang terkonsentrasi pada area lobar tertentu, khususnya pada bagian tengah dan bawah paru-paru, mengindikasikan bahwa model memfokuskan perhatian tertingginya pada area tersebut sebagai fitur utama dalam pengambilan keputusan klasifikasi pneumonia. Warna kuning dan hijau di sekitar area merah menunjukkan tingkat aktivasi sedang yang turut berkontribusi dalam proses klasifikasi, sedangkan warna biru pada bagian tepi citra merepresentasikan area dengan aktivasi terendah yang dianggap kurang relevan oleh model.

3.3.3 Normal

Pada Gambar 11(b) yang mempresentasikan citra paru-paru NORMAL, persebaran warna dari Grad-CAM terlihat lebih merata atau menyebar tipis (dominan biru-hijau) tanpa ada titik merah pekat yang mengunci area tertentu. Kondisi ini mengonfirmasi bahwa jaringan tidak mendeteksi adanya konsolidasi abnormal atau opasitas masif pada parenkim paru, yang merefleksikan kondisi radiologis paru-paru sehat di mana udara mengalir dengan baik tanpa hambatan cairan infeksi.



Gambar 12. Grad-CAM Normal

Berdasarkan gambar di atas, model FlexMatch-DenseNet169 berhasil mengklasifikasikan citra *Chest X-ray* sebagai kelas NORMAL dengan tingkat kepercayaan 99,66% dan status prediksi CORRECT.

Hasil ini menunjukkan bahwa prediksi model sesuai dengan *Ground Truth* yang telah ditetapkan.

Visualisasi *Grad-CAM Heatmap* memperlihatkan distribusi aktivasi model melalui gradasi warna yang merepresentasikan tingkat kepentingan setiap area pada citra. Warna merah terkonsentrasi pada area tengah hingga atas paru-paru, mengindikasikan bahwa model memfokuskan perhatian utamanya pada area tersebut untuk memverifikasi ketiadaan kelainan patologis. Warna kuning dan hijau di sekitar area merah menunjukkan tingkat aktivasi sedang, sementara warna biru pada bagian tepi citra merepresentasikan area dengan aktivasi terendah yang dianggap kurang relevan dalam proses klasifikasi.

3.4 Perbandingan dengan Penelitian Terdahulu

Tabel 4. Perbandingan dengan penelitian Terdahulu

Penelitian	Domain	Evaluasi	Capaian / Keterbatasan
Sohn et al.[34].(2020)	Klasifikasi citra (CIFAR-10) menggunakan FixMatch	Akurasi: 95,74%	Menggunakan ambang batas tetap; kurang optimal pada kelas dengan distribusi tidak seimbang
Zhang et al. [22](2021)	Klasifikasi citra semi-supervised menggunakan FlexMatch	Akurasi: 95,02%	Ambang batas dinamis per kelas; belum diujikan pada domain citra medis secara spesifik
Setiawan et al. [35](2022)	Klasifikasi Chest X-ray (COVID-19, Pneumonia, Normal) menggunakan DenseNet-121	Akurasi: 93,40%	Hanya menggunakan data berlabel; tidak memanfaatkan data tidak berlabel; tidak ada visualisasi Grad-CAM
Sukma Negara et al. [32] (2023)	Klasifikasi citra medis menggunakan Semi-Supervised Learning	Akurasi: 94,87%	ambang batas dinamis; evaluasi Grad-CAM terbatas

Berdasarkan tabel 4 perbandingan penelitian terdahulu, terlihat bahwa penelitian-penelitian sebelumnya memiliki keterbatasan yang signifikan dalam hal pemanfaatan data tidak berlabel dan fleksibilitas ambang batas klasifikasi. Sohn et al. (2020) dengan metode FixMatch hanya mampu mencapai akurasi sebesar 95,74% pada dataset CIFAR-10 dengan menggunakan ambang batas tetap, sehingga kurang optimal ketika diterapkan pada kelas dengan distribusi data yang tidak seimbang. Demikian pula, Zhang et al. (2021) yang mengembangkan FlexMatch dengan

ambang batas dinamis per kelas dan berhasil mencapai akurasi 95,02% dengan 40 label, namun belum diujikan secara spesifik pada domain citra medis. Sementara itu, Setiawan et al. (2022) yang mengklasifikasikan Chest X-ray menggunakan DenseNet-121 hanya memperoleh akurasi 93,40% karena seluruhnya bergantung pada data berlabel tanpa memanfaatkan data tidak berlabel, serta tidak dilengkapi dengan visualisasi Grad-CAM. Begitu pula Sukma Negara et al. (2023) yang meskipun menggunakan pendekatan *semi-supervised learning*, belum mengintegrasikan ambang batas dinamis secara penuh dan evaluasi Grad-CAM yang dihasilkan masih terbatas.

Dibandingkan dengan seluruh penelitian terdahulu tersebut, penelitian ini yang mengintegrasikan metode FlexMatch dengan arsitektur DenseNet-169 berhasil mencapai akurasi sebesar 96,65% pada klasifikasi tiga kelas Chest X-ray, yaitu COVID-19, Pneumonia, dan Normal. Capaian ini melampaui akurasi FixMatch (95,74%), FlexMatch standar (95,02%), DenseNet-121 berbasis *supervised* (93,40%), maupun pendekatan *semi-supervised* Sukma Negara et al. (94,87%). Keunggulan ini diperoleh melalui kombinasi mekanisme ambang batas dinamis per kelas yang memungkinkan pemanfaatan data berlabel dan tidak berlabel secara bersamaan, serta dilengkapi dengan visualisasi Grad-CAM sebagai bentuk validasi medis berbasis *Explainable AI* (XAI). Dengan demikian, penelitian ini tidak hanya unggul secara kuantitatif, tetapi juga memberikan kontribusi nyata dalam transparansi dan interpretabilitas model klasifikasi penyakit paru-paru.

4. Kesimpulan

Penelitian ini mampu menganalisis dan mengimplementasikan metode FlexMatch dalam kerangka *Semi-Supervised Deep Learning* untuk klasifikasi penyakit paru-paru berbasis citra Chest X-ray menggunakan arsitektur DenseNet-169. Dengan memanfaatkan mekanisme *Curriculum Pseudo Labeling* dan *class-adaptive thresholding*, model yang dikembangkan mampu mengoptimalkan pemanfaatan data berlabel maupun data tidak berlabel secara bersamaan. Hasil evaluasi menunjukkan bahwa model FlexMatch-DenseNet169 berhasil mencapai akurasi validasi sebesar 96,65%, dengan F1-Score masing-masing sebesar 99,02% untuk kelas COVID-19, 95,74% untuk kelas Normal, dan 94,48% untuk kelas Pneumonia. Capaian ini secara konsisten melampaui performa metode-metode terdahulu seperti FixMatch (95,74%), FlexMatch standar (95,02%), DenseNet-121 berbasis *supervised* (93,40%), maupun pendekatan *semi-supervised* lainnya (94,87%), sehingga membuktikan keunggulan nyata dari integrasi FlexMatch dengan arsitektur DenseNet-169. Selain itu, visualisasi Grad-CAM yang diterapkan pada ketiga kelas membuktikan bahwa model secara konsisten mampu memfokuskan perhatiannya pada area

patologis paru-paru yang relevan secara klinis, seperti *Ground-Glass Opacity* pada COVID-19, konsolidasi lobar pada pneumonia, serta distribusi aktivasi merata pada kondisi normal.

Meskipun penelitian ini mendapatkan performa yang memuaskan, terdapat beberapa keterbatasan yang perlu diperhatikan untuk pengembangan lebih lanjut. Pertama, dataset yang digunakan hanya mencakup 3.000 citra dari satu sumber repositori tunggal (*Mendeley Data*) dengan distribusi kelas yang seimbang, sehingga belum mencerminkan kondisi nyata di lingkungan klinis yang umumnya memiliki distribusi kelas tidak seimbang dan variasi kualitas citra yang lebih tinggi. Kedua, model hanya dievaluasi pada tiga kelas penyakit paru-paru, sehingga kemampuan generalisasinya terhadap penyakit paru-paru lain seperti *tuberkulosis*, *efusi pleura*, atau *atelektasis* belum dapat dipastikan. Ketiga, pelatihan dilakukan hanya selama 20 epoch dengan *hyperparameter* yang bersifat tetap, sehingga potensi optimasi lebih lanjut melalui teknik *fine-tuning* maupun pencarian *hyperparameter* adaptif masih terbuka lebar. Untuk penelitian selanjutnya, disarankan agar dilakukan pengujian model pada dataset yang lebih besar dan beragam dari berbagai fasilitas kesehatan, serta mengeksplorasi integrasi FlexMatch dengan arsitektur *backbone* yang lebih mutakhir seperti *Vision Transformer* (ViT) atau *EfficientNet*. Selain itu, penerapan teknik *federated learning* dapat menjadi alternatif yang menjanjikan untuk mengatasi keterbatasan ketersediaan data berlabel di lingkungan medis yang terdistribusi, sekaligus menjaga privasi data pasien secara lebih baik.

Daftar Rujukan

- [1] W. H. Organization, "Global Tuberculosis Report 2022," WHO, Geneva, 2022. [Online]. Available: <https://www.who.int/teams/global-programme-on-tuberculosis-and-lung-health/tb-reports/global-tuberculosis-report-2022>.
- [2] P. Rajpurkar, E. Chen, E. O'Keefe, and N. Shah, "Deep learning for chest radiograph diagnosis: A retrospective comparison of CheXNeXt to practicing radiologists," *PLoS Med.*, vol. 15, no. 11, p. e1002686, 2022, doi: <https://doi.org/10.1371/journal.pmed.1002686>.
- [3] J. A. Maharani and M. Masriadi, "Pengaruh Artificial Intellegence Dalam Mendeteksi Kasus Penyakit Tidak Menular," *J. Artif. Intell. Digit. Bus.*, vol. 4, no. 2, pp. 7343–7349, 2025, <https://doi.org/10.31004/riggs.v4i2.1864>.
- [4] C. J. Ryerson et al., "Update of the international multidisciplinary classification of the interstitial pneumonias: an ERS/ATS statement," *Eur. Respir. J.*, 2025, <https://doi.org/10.1183/13993003.00158-2025>.
- [5] M. Merk, "Still a role for the chest radiograph – humble but helpful!," *African J. Thorac. Crit. Care Med.*, vol. 28, p. 143, 2022, <https://doi.org/10.7196/AJTCCM.2022.v28i4.300>.
- [6] H. K. Ahmad et al., "Machine Learning Augmented Interpretation of Chest X-rays: A Systematic Review," *Diagnostics*, vol. 13, no. 4, p. 743, 2023.

- <https://doi.org/10.3390/diagnostics13040743>
- [7] P. Syifa, Safwandi, and Z. Fitri, "Sistem pakar diagnosis penyakit paru menggunakan metode convolutional neural network dan rule based system," *Rabit J. Teknol. Dan Sist. Inf. Univrab*, vol. 10, no. 2, pp. 1380–1392, 2025, <https://doi.org/10.1016/j.bea.2023.100076>
 - [8] S. Modak, "Applications of deep learning in disease diagnosis of chest radiographs: A survey on materials and methods," *Biomed. Eng. Adv.*, vol. 5, p. 100076, 2023, <https://doi.org/10.1016/j.bea.2023.100076>
 - [9] D. Banik and D. Bhattacharjee, "Mitigating Data Imbalance Issues in Medical Image Analysis," 2021, pp. 66–89. <https://doi.org/10.4018/978-1-7998-7371-6.ch004>
 - [10] Y. Xu, Y. Li, D. Wang, Y. Zhang, and D. Huang, "Addressing the current challenges in the clinical application of AI-based Radiomics for cancer imaging," *Front. Med.*, vol. 12, 2025, <https://doi.org/10.3389/fmed.2025.1674397>
 - [11] A. Chowdhury, J. Rosenthal, J. Waring, and R. Umeton, "Applying Self-Supervised Learning to Medicine: Review of the State of the Art and Medical Implementations," *Informatics*, vol. 8, no. 3, p. 59, 2021, <https://doi.org/10.3390/informatics8030059>
 - [12] E. M. J. S. Fdo, A. A. A. Rekshin, G. G. Gains, X. A. M. Mcchenzi, and A. J. P. Peniel, "Self-supervised learning for small-scale medical imaging dataset," *World J. Adv. Eng. Technol. Sci.*, vol. 13, no. 2, pp. 142–145, 2024, <https://doi.org/10.30574/wjaets.2024.13.2.0526>
 - [13] L.-Z. Guo, L.-H. Jia, J.-J. Shao, and Y. Li, "Robust semi-supervised learning in open environments," *Front. Comput. Sci.*, vol. 19, no. 8, 2025, <https://doi.org/10.1007/s11704-024-40646-w>
 - [14] B. Zhang *et al.*, "FlexMatch: Boosting Semi-Supervised Learning with Curriculum Pseudo Labeling," *N/A*, p. 34, 2021, [Online]. Available: <https://proceedings.neurips.cc/paper/2021/hash/995693c15f439e3d189b06e89d145dd5-Abstract.html>
 - [15] G. Gui, Z. Zhao, L. Qi, L. Zhou, L. Wang, and Y.-H. Shi, "Enhancing Sample Utilization through Sample Adaptive Augmentation in Semi-Supervised Learning," *arXiv.Org*, 2023, <https://doi.org/10.1109/ICCV51070.2023.01455>
 - [16] V. Leon, A. Pasko, and B. Terenzio, "Progress and Prospects in Deep Learning for Chest X-Ray Interpretation," 2025, <https://doi.org/10.36227/techrxiv.173592311.10695741/v1>
 - [17] M. O. Sahetai, D. Chauhan, and R. Jani, "Harnessing Artificial Intelligence for Precise Pulmonary Disease Diagnosis," 2023, pp. 151–157, <https://doi.org/10.1109/ICSSAS57918.2023.10331667>
 - [18] S. Kumar, "Covid19-Pneumonia-Normal Chest X-Ray Images," vol. VI. Mendeley Data, 2022, doi: 10.17632/dvntn9yhd2.1.
 - [19] 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, <https://doi.org/10.3390/diagnostics12061457>
 - [20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778, <https://doi.org/10.1109/CVPR.2016.90>
 - [21] X. Zhao, S. Qi, B. Zhang, H. Lian, S. Yao, and W. Sun, "Deep Learning for Pulmonary Nodule Detection: A Review," *Front. Oncol.*, vol. 12, p. 837279, 2022, <https://doi.org/10.3389/fonc.2022.983833>
 - [22] 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, <https://doi.org/10.3389/fmed.2021.629134>
 - [23] H. Jin, H. Che, Y. Lin, and H. Chen, "Promptmrg: Diagnosis-driven prompts for medical report generation," 2024, [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/download/28038/28087>
 - [24] Y. Wang *et al.*, "FlexMatch: Boosting Semi-Supervised Learning with Curriculum Pseudo Labeling," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 25324–25336, 2021.
 - [25] 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, <https://doi.org/10.1007/s42979-022-01627-7>
 - [26] M. N. Dawami, B. S. Negara, M. Irsyad, and F. Yanto, "Klasifikasi Multikelas Citra Chest X-Ray Menggunakan Semi-Supervised SoftMatch pada Label Terbatas," vol. 6, no. 12, pp. 2345–2360, 2026, doi: 10.47065/tin.v6i12.9848.
 - [27] M. Shetty and S. Shetty, "Computer-Aided Diagnostic of COVID-19 Using Chest X-Ray Analysis BT - Proceedings of Second Doctoral Symposium on Computational Intelligence," 2022, pp. 675–681. https://doi.org/10.1007/978-981-16-3346-1_54
 - [28] G. Huang and K. Q. Weinberger, "Densely Connected Convolutional Networks," doi: <https://doi.org/10.48550/arXiv.1608.06993>.
 - [29] F. Shahira and B. S. Negara, "Lung X-Ray Image Classification Using DenseNet-169 and Bayesian Optimization," vol. 05, no. 01, pp. 18–27, 2025, <https://doi.org/10.30983/knownbase.v5i1.9618>
 - [30] J. Van Zyl, "Analysis of classification metric behaviour under class imbalance," *Egypt. Informatics J.*, vol. 31, no. July, p. 100711, 2025, doi: 10.1016/j.eij.2025.100711.
 - [31] E. Selby, M. Woodruff, and J. Carter, "Explainable AI in Medical Imaging: A Systematic Review of Grad-CAM Applications," *J. Med. Imaging Heal. Informatics*, vol. 13, no. 2, pp. 145–162, 2023, doi: 10.1166/jmhi.2023.3821.
 - [32] B. Sukma Negara, T. Hariyanto, and R. Pratiwi, "Grad-CAM Visualization on DenseNet Architecture for Chest X-Ray Classification," *Int. J. Intell. Syst. Appl. Eng.*, vol. 11, no. 3, pp. 512–521, 2023, doi: 10.18201/ijisae.2023.412.
 - [33] E. Tjoa and C. Guan, "A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 32, no. 11, pp. 4793–4813, 2021, <https://doi.org/10.1109/TNNLS.2020.3027314>
 - [34] K. Sohn *et al.*, "FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020, vol. 33, doi: 10.48550/arXiv.2001.07685.
 - [35] A. Setiawan, I. K. E. Purnama, and M. H. Purnomo, "Chest X-Ray Classification for COVID-19 Detection Using DenseNet-121," *J. Electron. Electromed. Eng. Med. Informatics*, vol. 4, no. 2, pp. 78–86, 2022, <https://doi.org/10.35882/jeeemi.v4i2.5>