

**DETEKSI PNEUMONIA PADA CITRA X-RAY MENGGUNAKAN *CNN*
DENGAN *STACKING ENSEMBLE***

SKRIPSI

Oleh:

RIDHO AULIA RAHMAN

NIM. 220605110044



**PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2025**

**DETEKSI PNEUMONIA PADA CITRA X-RAY MENGGUNAKAN *CNN*
DENGAN *STACKING ENSEMBLE***

SKRIPSI

Diajukan kepada:

Universitas Islam Negeri Maulana Malik Ibrahim Malang
Untuk memenuhi Salah Satu Persyaratan dalam
Memperoleh Gelar Sarjana Komputer (S.Kom)

Oleh :

RIDHO AULIA RAHMAN
NIM. 220605110044

**PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2025**

HALAMAN PERSETUJUAN

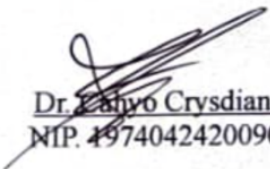
DETEKSI PNEUMONIA PADA CITRA X-RAY MENGGUNAKAN CNN DENGAN STACKING ENSEMBLE

SKRIPSI

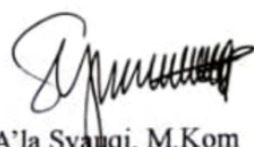
Oleh :
RIDHO AULIA RAHMAN
NIM. 220605110052

Telah Diperiksa dan Disetujui untuk Diuji:
Tanggal: 11 Desember 2025

Pembimbing I,


Dr. Zuhro Crysdian, M.Cs
NIP. 197404242009011008

Pembimbing II,


A'la Syauqi, M.Kom
NIP. 19771201 200801 1 007

Mengetahui,
Ketua Program Studi Teknik Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang



Supriyono, M. Kom
NIP. 19841010 201903 1 012

HALAMAN PENGESAHAN

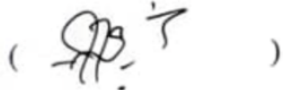
DETEKSI PNEUMONIA PADA CITRA X-RAY MENGGUNAKAN CNN DENGAN *STACKING ENSEMBLE*

SKRIPSI

Oleh :
RIDHO AULIA RAHMAN
NIM. 220605110044

Telah Dipertahankan di Depan Dewan Penguji Skripsi
dan Dinyatakan Diterima Sebagai Salah Satu Persyaratan
Untuk Memperoleh Gelar Sarjana Komputer (S.Kom)
Tanggal: 1 Desember 2025

Susunan Dewan Penguji

Ketua Penguji	: <u>Dr. Zainal Abidin, M.Kom</u> NIP. 19760613 200501 1 004	()
Anggota Penguji I	: <u>Khadijah Fahmi Hayati Holle, M.Kom</u> NIP. 19900626 202203 2 002	()
Anggota Penguji II	: <u>Dr. Cahyo Crysdian, M.Cs</u> NIP. 197404242009011008	()
Anggota Penguji III	: <u>A'la Syaqui, M.Kom</u> NIP. 19771201 200801 1 007	()

Mengetahui dan Mengesahkan,
Ketua Program Studi Teknik Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang




Supriyono, M. Kom
NIP. 19841010 201903 1 012

PERNYATAAN KEASLIAN TULISAN

Saya yang bertanda tangan di bawah ini:

Nama : Ridho Aulia Rahman
NIM : 220605110044
Fakultas /Program Studi : Sains dan Teknologi / Teknik Informatika
Judul Skripsi : DETEKSI PNEUMONIA PADA CITRA X-RAY
MENGGUNAKAN *CNN* DENGAN *STACKING*
ENSEMBLE.

Menyatakan dengan sebenarnya bahwa Skripsi yang saya tulis ini benar-benar merupakan hasil karya saya sendiri, bukan merupakan pengambil alihan data, tulisan, atau pikiran orang lain yang saya akui sebagai hasil tulisan atau pikiran saya sendiri, kecuali dengan mencantumkan sumber cuplikan pada daftar pustaka.

Apabila dikemudian hari terbukti atau dapat dibuktikan skripsi ini merupakan hasil jiplakan, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Malang, 11 Desember 2025
Yang membuat pernyataan,



Ridho Aulia Rahman
NIM. 220605110044

MOTTO

“Dunia hanya sementara , akhirat selamanya”

“Do it now or never.”

“fastabiqul Khoirot, Khoirunnas anfa’uhum lin nass”

HALAMAN PERSEMBAHAN

Alhamdulillah rabbil alamin saya ucapkan atas hasil skripsi yang telah tertulishingga tuntas. Untuk itu, saya persembahkan skripsi ini kepada:

- Kedua orang tua saya, yakni Eko Andrihasno S.E dan Bdn. AMIRAH DAULATIR RACHMAH, STr.Keb. yang telah membimbing, mendoakan, dan menyayangi saya dari awal saya berada di dunia hingga saat ini.
- Pengurus Takmir Masjid Nurun Nuha, yang mana telah memberikan wejangan dan nasihat luar biasa atas kehidupan ini.
- Dosen saya, Bapak Dr. Cahyo Crysdiyan, M.Cs., Dr. Zainal Abidin M.Kom, A'la Syauqi M.Kom, Ibu Khadijah Fahmi Hayati Holle, M.Kom, yang telah mengajari dan membukakan pintu wawasan yang sangat luas mengenai dunia riset dan akademik kepada saya.
- Untuk teman seperjuangan Infinity 2022 khususnya S, terimakasih dukungan, diskusi , dan semua hal baik kita semua.
- Dan terakhir, untuk diriku sendiri, yang telah bertahan, belajar, dan berjuang melewati setiap tantangan hingga akhirnya karya ini dapat terselesaikan.

Semoga kita semua selalu mendapat petunjuk mengenai jalan yang lurus dari Allah serta semoga kita dapat mewujudkan harapan dan cita – cita yang telah diimpikan.Amin Allahumma Amin.

KATA PENGANTAR

Segala puji hanya milik Allah Subhanahu Wa Ta'ala atas segala nikmat dankasih sayang-Nya yang telah memudahkan penulis untuk menyelesaikan skripsi yang berjudul “DETEKSI PNEUMONIA PADA CITRA X-RAY MENGGUNAKAN CNN DENGAN STACKING ENSEMBLE”. Semoga shalawat dan salam senantiasa terlimpah kepada Nabi Muhammad Sallallahu ‘Alaihi wa Sallam. Dan semoga kita semua mendapat syafaatnya di hari kiamat nanti, Aamiin.

Penulis mengucapkan rasa syukur dan terima kasih yang tak terhingga kepada semua pihak-pihak yang selalu memberikan bantuan dan motivasi kepada penulis untuk dapat menyelesaikan skripsi ini. Ucapan ini penulis sampaikan kepada:

1. Prof. Dr. Hj. Ilfi Nur Diana, M.Si, selaku Rektor Universitas Islam Negeri Malang, yang telah memberikan kesempatan dan fasilitas untuk menempuh pendidikan di universitas ini.
2. Dr. Agus Mulyono, M.Kes, selaku Dekan Fakultas Saintek Universitas Islam Negeri Malang, atas segala dukungan dan arahan yang diberikan selama proses studi.
3. Dr. Cahyo Crysdiyan, M.Cs. selaku dosen pembimbing I dan A'la Syauqi M.Kom. selaku dosen pembimbing II yang telah memberikan bantuan dan arahan kepada penulis, sehingga bisa menuntaskan skripsi ini.
4. Dr. Zainal Abidin M.Kom. selaku dosen penguji I dan Khadijah Fahmi Hayati Holle, M.Kom selaku dosen penguji II yang telah menguji

serta memberikan masukan sehingga penulis dapat menuntaskan skripsi dengan baik.

5. Seluruh dosen dan staf Fakultas Saintek Universitas Islam Negeri Malang, atas ilmu, bimbingan, dan bantuan administratif yang telah diberikan selama penulis menempuh studi.

Akhir kata, penulis mengakui bahwa penulisan pada skripsi ini masih banyak kekurangan. Saya berharap semoga skripsi ini diterima sebagai amal ibadah yang tulus dan bermanfaat di sisi Allah Subhanahu Wa Ta'ala. Semoga karya ini menjadi bagian dari kontribusi yang tak terputus dalam rangka memperkuat dan mengembangkan ilmu pengetahuan, serta melaksanakan tugas sebagai hamba Allah yang berkomitmen. Wassalamualaikum Warahmatullahi Wabarakatuh.

Malang, 11 Desember 2025

Penulis

DAFTAR ISI

HALAMAN PENGAJUAN	ii
HALAMAN PERSETUJUAN.....	iii
HALAMAN PENGESAHAN	iv
PERNYATAAN KEASLIAN TULISAN	v
MOTTO	vi
HALAMAN PERSEMBAHAN.....	vii
KATA PENGANTAR.....	viii
DAFTAR ISI.....	x
DAFTAR GAMBAR.....	xii
DAFTAR TABEL.....	xiv
ABSTRAK	xv
ABSTRACT	xvi
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang.....	1
1.2 Pernyataan Masalah	4
1.3 Tujuan Penelitian	4
1.4 Manfaat Penelitian	4
1.5 Batasan Masalah	5
BAB II STUDI PUSTAKA	6
BAB III DESAIN DAN IMPLEMENTASI SISTEM KLASIFIKASI	
PNEUMONIA	19
3.1 Prosedur Penelitian	19
3.2 Persiapan Data	20
3.3 Desain Sistem	23
3.3.1 Input Image	23
3.3.2 Preprocessing Citra	24
3.3.3 Ekstraksi Fitur	27
3.3.4 Classifier.....	30
3.3.5 Stacking Ensemble	34
3.4 Implementasi Sistem.....	41
BAB IV UJI COBA DAN PEMBAHASAN.....	42
4.1 Skenario Pengujian	42
4.2 Hasil Pengujian.....	53
4.2.1 Hasil Pengujian Model Tunggal: EfficientNetB0	53
4.2.2 Hasil Pengujian Model Tunggal: MobileNetV2	66
4.2.3 Hasil Pengujian Model Tunggal: MobileNetV3Small	78
4.2.4 Hasil Pengujian Model Stacking Ensemble	90
4.3 Analisis dan Pembahasan	107
4.3.1 Analisis Performa Model pada Berbagai Skenario Pengujian.....	108
4.3.2 Perbandingan Stacking Ensemble dan Model Tunggal Berdasarkan Uji Statistik (t-test)	114

4.3.3	Analisis Pengaruh Penerapan SMOTE terhadap Performa Model Berdasarkan Uji t-Test	116
4.3.4	Implikasi Hasil Penelitian terhadap Aplikasi Klinis dan Integrasi Nilai Keislaman.....	118
BAB V KESIMPULAN DAN SARAN		120
5.1	Kesimpulan.....	120
5.2	Saran.....	120
DARTAR PUSTAKA		

DAFTAR GAMBAR

Gambar 3. 1 Desain Penelitian.....	20
Gambar 3. 2 Desain Sistem.....	23
Gambar 3. 3 <i>Preprocessing</i> Citra.....	24
Gambar 3. 4 Normalisasi Nilai Piksel.....	26
Gambar 3. 5 Ekstraksi Fitur <i>Pre Trained Model</i>	28
Gambar 3. 6 Flowchart ekstraksi fitur.....	30
Gambar 3. 7 Arsitektur MobileNetV2 (Sandler et al., 2018).....	31
Gambar 3. 8 Arsitektur MobileNetV3-Small (Howard et al., 2019)	32
Gambar 3. 9 Arsitektur EfficientNet-B0 (Tan & Le, 2019).....	32
Gambar 3. 10 <i>Classifier (Base Model)</i>	34
Gambar 3. 11 Arsitektur Stacking Ensemble	36
Gambar 3. 12 Arsitektur <i>Meta Learner (Dense Neural Network)</i>	39
Gambar 4. 1 Flowchart Pengujian.....	45
Gambar 4. 2 ROC Curve EfficientNetB0 (Skenario 1A-K)	56
Gambar 4. 3 Confusion Matrix EfficientNetB0 (Skenario 1A-K).....	56
Gambar 4. 4 Precision-Recall Curve EfficientNetB0 (Skenario 1A-K)	57
Gambar 4. 5 ROC Curve EfficientNetB0 (Skenario 1B-K).....	58
Gambar 4. 6 Precision-Recall Curve EfficientNetB0 (Skenario 1B-K)	59
Gambar 4. 7 Confusion Matrix EfficientNetB0 (Skenario 1B-K)	59
Gambar 4. 8 ROC Curve EfficientNetB0 (Skenario 1A-F)	61
Gambar 4. 9 Confusion Matrix EfficientNetB0 (Skenario 1A-F).....	62
Gambar 4. 10 Precision-Recall Curve EfficientNetB0 (Skenario 1A-F).....	62
Gambar 4. 11 ROC Curve EfficientNetB0 (Skenario 1B-F)	64
Gambar 4. 12 Confusion Matrix EfficientNetB0 (Skenario 1B-F).....	65
Gambar 4. 13 Precision-Recall Curve EfficientNetB0 (Skenario 1B-F)	65
Gambar 4. 14 ROC Curve MobileNetV2 (Skenario 2A-K).....	68
Gambar 4. 15 Precision-Recall Curve MobileNetV2 (Skenario 2A-K)	68
Gambar 4. 16 Confusion Matrix MobileNetV2 (Skenario 2A-K)	69
Gambar 4. 17 ROC Curve MobileNetV2 (Skenario 2B-K).....	71
Gambar 4. 18 Precision-Recall Curve MobileNetV2 (Skenario 2B-K).....	71
Gambar 4. 19 Confusion Matrix MobileNetV2 (Skenario 2B-K)	72
Gambar 4. 20 ROC Curve MobileNetV2 (Skenario 2A-F)	74
Gambar 4. 21 Precision–Recall Curve MobileNetV2 (Skenario 2A-F)	74
Gambar 4. 22 Confusion Matrix MobileNetV2 (Skenario 2A-F).....	75
Gambar 4. 23 ROC Curve MobileNetV2 (Skenario 2B-F).....	77
Gambar 4. 24 Precision–Recall Curve MobileNetV2 (Skenario 2B-F).....	77
Gambar 4. 25 Confusion Matrix MobileNetV2 (Skenario 2B-F).....	78
Gambar 4. 26 ROC Curve MobileNetV3Small (Skenario 3A-K)	80
Gambar 4. 27 Precision-Recall Curve MobileNetV3Small (Skenario 3A-K)	81
Gambar 4. 28 Confusion Matrix MobileNetV3Small (Skenario 3A-K).....	81
Gambar 4. 29 ROC Curve MobileNetV3Small (Skenario 3B-K).....	83
Gambar 4. 30 Precision-Recall Curve MobileNetV3Small (Skenario 3B-K)	84
Gambar 4. 31 Confusion Matrix MobileNetV3Small (Skenario 3B-K).....	84

Gambar 4. 32 ROC Curve MobileNetV3Small (Skenario 3A-F)	86
Gambar 4. 33 Precision–Recall Curve MobileNetV3Small (Skenario 3A-F)	86
Gambar 4. 34 Confusion Matrix MobileNetV3Small (Skenario 3A-F)	87
Gambar 4. 35 ROC Curve MobileNetV3Small (Skenario 3B-F)	89
Gambar 4. 36 Precision–Recall Curve MobileNetV3Small (Skenario 3B-F)	89
Gambar 4. 37 Confusion Matrix MobileNetV3Small (Skenario 3B-F).....	90
Gambar 4. 38 ROC Curve Stacking Ensemble (Skenario 4A-K)	92
Gambar 4. 39 Precision-Recall Curve Stacking Ensemble (Skenario 4A-K)	93
Gambar 4. 40 Confusion Matrix Stacking Ensemble (Skenario 4A-K).....	93
Gambar 4. 41 ROC Curve Stacking Ensemble (Skenario 4B-K)	95
Gambar 4. 42 Precision-Recall Curve Stacking Ensemble (Skenario 4B-K)	95
Gambar 4. 43 Confusion Matrix Stacking Ensemble (Skenario 4B-K).....	96
Gambar 4. 44 ROC Curve Stacking Ensemble (Skenario 4A-F).....	98
Gambar 4. 45 Precision–Recall Curve Stacking Ensemble (Skenario 4A-F).....	98
Gambar 4. 46 Confusion Matrix Stacking Ensemble (Skenario 4A-F)	99
Gambar 4. 47 Hasil Validasi Tiap Fold Base Models pada Skenario 4B-F.....	99
Gambar 4. 48 ROC Curve Stacking Ensemble (Skenario 4B-F)	100
Gambar 4. 49 Precision–Recall Curve Stacking Ensemble (Skenario 4B-F)	101
Gambar 4. 50 Confusion Matrix Stacking Ensemble (Skenario 4B-F)	101
Gambar 4. 51 Confusion Matrix EfficientNetB0 dengan SMOTE.....	103
Gambar 4. 52 AUC EfficientNetB0 dengan SMOTE	103
Gambar 4. 53 ROC EfficientNetB0 dengan SMOTE	104
Gambar 4. 54 Confusion Matrix Stacking Ensemble dengan SMOTE	105
Gambar 4. 55 AUC Stacking Ensemble dengan SMOTE.....	106
Gambar 4. 56 ROC Stacking Ensemble dengan SMOTE.....	106
Gambar 4. 57 Perbandingan Akurasi Antar Skenario.....	112
Gambar 4. 58 Perbandingan AUC-ROC Antar Skenario.....	113
Gambar 4. 59 Perbandingan F1-Score Antar Skenario.....	113
Gambar 4. 60 Perbandingan <i>Precision</i> Antar Skenario	114
Gambar 4. 61 Perbandingan Sensitivitas Antar Skenario	114

DAFTAR TABEL

Tabel 2. 1 Rangkuman Performa Model Deteksi Pneumonia.....	16
Tabel 3. 1 Sebaran dataset.....	21
Tabel 3. 2 Contoh gambar dataset citra X-ray	22
Tabel 4. 1 Kombinasi Skenario Pengujian	48
Tabel 4. 2 Nilai masing-masing parameter	50
Tabel 4. 3 Hasil Evaluasi Model EfficientNetB0.....	54
Tabel 4. 4 Log Training Skenario 1A-K Fold = 5	54
Tabel 4. 5 Log Training Skenario 1B-K Fold = 5.....	57
Tabel 4. 6 Log Training Skenario 1A-F Fold = 5	60
Tabel 4. 7 Log Training Skenario 1B-F Fold = 5	63
Tabel 4. 8 Hasil Evaluasi Model MobileNetV2.....	66
Tabel 4. 9 Log Training Skenario 2A-K Fold = 5	66
Tabel 4. 10 Log Training Skenario 2B-K Fold = 5.....	69
Tabel 4. 11 Log Training Skenario 2A-F Fold = 5	72
Tabel 4. 12 Log Training Skenario 2B-F Fold = 5	75
Tabel 4. 13 Hasil Evaluasi Model MobileNetV3Small	79
Tabel 4. 14 Log Training Skenario 3A-K Fold = 5	79
Tabel 4. 15 Log Training Skenario 3B-K Fold = 5.....	82
Tabel 4. 16 Log Training Skenario 3A-F Fold = 5	85
Tabel 4. 17 Log Training Skenario 3B-F Fold = 5	87
Tabel 4. 18 Hasil Evaluasi Model Stacking Ensemble	91
Tabel 4. 19 Hasil Validasi Tiap Fold Base Models pada Skenario 4A-K.....	91
Tabel 4. 20 Log Hasil Training Skenario 4B-K (Stacking Ensemble, 70:30, K=5)	94
Tabel 4. 21 Hasil Validasi Tiap Fold Base Models pada Skenario 4A-F	96
Tabel 4. 22 Hasil Evaluasi Performa Seluruh Model.....	109

ABSTRAK

Rahman, Ridho Aulia. 2025. **Deteksi Pneumonia pada Citra X-Ray Paru-Paru Menggunakan Pendekatan Stacking Ensemble**. Skripsi. Jurusan Teknik Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang. Pembimbing: (I) Dr. Cahyo Crsydian, M.Cs (II) A'la Syauqi, M.Kom.

Pneumonia merupakan infeksi paru yang berpotensi fatal apabila tidak terdeteksi secara dini, sehingga diperlukan sistem diagnosis berbantuan komputer yang akurat dan stabil. Penelitian ini bertujuan mengevaluasi apakah pendekatan *stacking ensemble* Convolutional Neural Network (CNN) mampu memberikan performa yang lebih baik dibandingkan model tunggal dalam deteksi pneumonia pada citra X-ray paru-paru. Metode yang diusulkan mengombinasikan EfficientNetB0, MobileNetV2, dan MobileNetV3Small sebagai *base learner* dengan Dense Neural Network (DNN) sebagai *meta-learner*, serta dievaluasi menggunakan metrik akurasi, sensitivitas, spesifisitas, F1-score, dan AUC-ROC pada berbagai skenario pembagian data dan ketidakseimbangan kelas. Hasil eksperimen menunjukkan bahwa model stacking ensemble CNN memberikan performa paling unggul dan konsisten dibandingkan seluruh model tunggal, dengan capaian terbaik pada skenario *full dataset* rasio 80:20 berupa akurasi 99,14%, sensitivitas 98,99%, dan AUC-ROC 99,87%, serta stabilitas yang lebih baik terhadap variasi data. Temuan ini membuktikan bahwa pendekatan stacking ensemble CNN lebih efektif dan andal untuk diterapkan sebagai sistem pendukung skrining klinis pneumonia.

Kata kunci: Pneumonia, Citra X-Ray, *Deep Learning*, *Stacking Ensemble*, CNN, *Dense Neural Network*

ABSTRACT

Rahman, Ridho 2025. **Pneumonia Detection on Chest X-Ray Images Using a Stacking Ensemble Approach**. Undergraduate Thesis. Department of Informatics Engineering, Faculty of Science and Technology, State Islamic University of Maulana Malik Ibrahim Malang. Advisors: (I) Dr. Cahyo Crsydian, M.Cs. (II) A'la Syauqi, M.Kom.

Pneumonia is a lung infection that can be fatal if not detected early, highlighting the need for accurate and stable computer-aided diagnosis systems. This study aims to evaluate whether a convolutional neural network (CNN) stacking ensemble approach can achieve better performance than single models in detecting pneumonia from chest X-ray images. The proposed method combines EfficientNetB0, MobileNetV2, and MobileNetV3Small as base learners with a Dense Neural Network (DNN) as a meta-learner, and is evaluated using accuracy, sensitivity, specificity, F1-score, and AUC-ROC across various data-splitting scenarios and class-imbalance conditions. Experimental results show that the CNN stacking ensemble consistently outperforms all single models, achieving its best performance on the full dataset with an 80:20 split, yielding 99.14% accuracy, 98.99% sensitivity, and an AUC-ROC of 99.87%, while demonstrating greater stability against data variability. These findings confirm that the CNN stacking ensemble is a more effective and reliable approach for implementation in computer-aided clinical pneumonia screening systems.

Keywords: Pneumonia, Chest X-Ray, Deep Learning, Stacking Ensemble, CNN, Dense Neural Network

البحث مستخلص

رحمان، رِضْهٌ أُولَيَا. ٢٠٢٥. اكتشاف الالتهاب الرئوي في صور الأشعة السينية للصدر باستخدام نهج التكديس التجميعي. رسالة البكالوريوس قسم هندسة المعلوماتية، كلية العلوم والتكنولوجيا، الجامعة الإسلامية الحكومية مولانا مالك إبراهيم مالانج. المشرفان: (١) الدكتور كاهيو كريسديان، ماجستير في علوم الحاسوب. (٢) علاء شوقي، ماجستير في علوم الحاسوب.

الكلمات المفتاحية: الالتهاب الرئوي، الأشعة السينية، التعلم العميق، التكديس التجميعي، الشبكات العصبية الكثيفة

يُعد الالتهاب الرئوي من الأمراض المعدية الخطيرة التي قد تكون قاتلة إذا لم يتم تشخيصها مبكرًا. يهدف هذا البحث إلى تطوير نموذج ذكاء اصطناعي للكشف التلقائي عن الالتهاب الرئوي اعتمادًا على صور الأشعة السينية للصدر باستخدام تقنيات التعلم العميق. استخدم الباحث نهج التكديس التجميعي (Stacking Ensemble) الذي يجمع بين ثلاثة نماذج أساسية: EfficientNetB0، MobileNetV2 و MobileNetV3Small، مع متعلم فوقي (Meta-Learner) يعتمد على شبكة عصبية كثيفة (Dense Neural Network). أظهرت النتائج أن نموذج التكديس التجميعي حقق دقة تتراوح بين 96.67% و 97.25%، ودقة تنبؤ (Precision) بلغت 98.14%، وتخصص (Specificity) 97.80%، ومنحنى ROC-AUC قريب من الكمال (0.9965–0.9970) بينما حقق النموذج الفردي الأفضل، EfficientNetB0، أعلى حساسية بلغت 98.62%. تشير النتائج إلى أن التكديس لا يتفوق تمامًا على النموذج الفردي الأفضل من حيث الدقة، لكنه يقلل بشكل ملحوظ من الإيجابيات الكاذبة. لذلك، يمكن اعتماد نهج التكديس في أنظمة دعم القرار الطبي، خاصة في مرحلة تأكيد تشخيص الالتهاب الرئوي.

BAB I

PENDAHULUAN

1.1 Latar Belakang

Pneumonia adalah penyakit infeksi saluran pernapasan akut yang menjadi salah satu penyebab utama kematian di seluruh dunia, terutama pada anak-anak di bawah usia lima tahun dan lansia. Data dari Organisasi Kesehatan Dunia (WHO) menyebutkan bahwa pneumonia bertanggung jawab atas sekitar 15% kematian anak balita secara global (WHO, 2020). Penyakit ini menyerang paru-paru, menyebabkan peradangan yang dapat mengganggu fungsi pernapasan, sehingga diagnosis dini dan tepat menjadi kunci untuk meningkatkan peluang kesembuhan dan mencegah dampak yang lebih serius. Salah satu cara utama untuk mendiagnosis pneumonia adalah melalui citra X-ray paru-paru, yang mampu menunjukkan kelainan seperti infiltrasi atau konsolidasi, tanda khas pneumonia (Cherian et al., 2005).

Namun, membaca citra X-ray secara manual oleh radiolog bukanlah hal yang mudah. Ada tantangan seperti perbedaan interpretasi antar dokter, keterbatasan waktu, hingga risiko kesalahan, terutama di fasilitas kesehatan dengan sumber daya terbatas (Kallio et al., 2019). Di sinilah teknologi kecerdasan buatan, khususnya pembelajaran mesin dan pembelajaran mendalam (*deep learning*), menawarkan harapan baru. Pendekatan berbasis *Convolutional Neural Networks* (CNN) telah terbukti mampu mengenali pola kelainan pada citra X-ray dengan akurasi yang menjanjikan (Rajpurkar et al., 2017)

Meski begitu, satu model saja sering kali kesulitan menangani keragaman data citra, seperti perbedaan kualitas gambar akibat variasi peralatan radiografi, kondisi pencahayaan, hingga noise pada proses akuisisi citra. Selain itu, variasi klinis dari pneumonia misalnya perbedaan antara pneumonia bakterial dan viral sering menimbulkan pola radiologis yang tidak seragam, sehingga model tunggal rawan mengalami overfitting atau bias terhadap distribusi data tertentu (Lakhani & Sundaram, 2017).

Penelitian lain menunjukkan bahwa CNN tunggal cenderung memiliki performa yang menurun ketika diuji pada data dari rumah sakit atau populasi yang berbeda, menandakan keterbatasan generalisasi (Zech et al., 2018). Faktor-faktor ini menjadikan penggunaan pendekatan tunggal kurang optimal dalam praktik klinis.

Untuk mengatasi keterbatasan tersebut, penelitian ini menerapkan pendekatan *stacking ensemble* yang mengombinasikan beberapa arsitektur CNN. Berbeda dengan *bagging* atau *boosting* yang berfokus pada pengurangan variansi dan bias melalui teknik pengambilan sampel atau penyesuaian bobot (Breiman, 1996; Freund & Schapire, 1997), *stacking* memanfaatkan keberagaman model dengan mengintegrasikan prediksi dari beberapa *base learners* menggunakan *meta-learner*. Pendekatan ini terbukti mampu meningkatkan kestabilan keputusan dan akurasi pada tugas klasifikasi citra medis (Wolpert, 1992; Zhou, 2012). Dengan demikian, penggunaan *stacking ensemble* menjadi lebih relevan untuk memastikan sistem deteksi pneumonia dapat bekerja lebih akurat dan andal di berbagai konteks klinis.

Dari perspektif Islam, menjaga kesehatan adalah bagian dari amanah yang diberikan Allah kepada manusia. Al-Qur'an dalam Surah Al-Baqarah ayat 195

﴿١٩٥﴾ وَأَنْفِقُوا فِي سَبِيلِ اللَّهِ وَلَا تُلْقُوا بِأَيْدِيكُمْ إِلَى التَّهْلُكَةِ وَأَحْسِنُوا إِنَّ اللَّهَ يُحِبُّ الْمُحْسِنِينَ

“Berinfaklah di jalan Allah, janganlah jerumuskan dirimu ke dalam kebinasaan, dan berbuatbaiklah. Sesungguhnya Allah menyukai orang-orang yang berbuat baik.” (QS. Al-Baqarah-195)

menyebutkan, “Dan janganlah kamu menjatuhkan dirimu sendiri ke dalam kebinasaan.” Menyediakan alat bantu diagnosis yang akurat seperti sistem berbasis *stacking ensemble* ini sejalan dengan prinsip menjaga nyawa dan mempercepat penyembuhan, yang merupakan wujud dari *rahmatan lil ‘alamin*. Dengan memanfaatkan teknologi, kita dapat mendukung upaya menyelamatkan jiwa, terutama di komunitas yang kekurangan akses ke fasilitas kesehatan memadai, sehingga mencerminkan nilai kasih sayang dan keadilan dalam Islam.

Berdasarkan latar belakang ini, penelitian ini bertujuan untuk mengembangkan sistem deteksi pneumonia pada citra X-ray paru-paru menggunakan pendekatan *stacking ensemble*. Sistem ini diharapkan dapat menjadi solusi yang lebih akurat dan andal, mendukung tenaga medis dalam mendiagnosis pneumonia, khususnya di wilayah dengan keterbatasan sumber daya, sekaligus mewujudkan nilai-nilai kemanusiaan yang selaras dengan ajaran Islam.

1.2 Pernyataan Masalah

Apakah model *stacking ensemble* CNN memiliki performa yang lebih baik daripada model tunggal dalam deteksi pneumonia pada citra X-ray berdasarkan akurasi, sensitivitas, spesifisitas, F1-score, dan AUC-ROC?

1.3 Tujuan Penelitian

Untuk mengukur performa model *stacking ensemble* dengan model pembelajaran mesin tunggal dalam mendeteksi pneumonia berbasis citra X-ray berdasarkan akurasi, sensitivitas, spesifisitas, F1-score, dan AUC-ROC.

1.4 Manfaat Penelitian

Dengan dilakukannya penelitian ini diharapkan dapat memberikan manfaat di masa depan, yaitu:

1. Hasil penelitian ini dapat diimplementasikan di puskesmas maupun rumah sakit daerah, termasuk wilayah 3T (Tertinggal, Terdepan, Terluar), sebagai alat bantu deteksi pneumonia yang mendukung diagnosis cepat dan akurat.
2. Bagi Kementerian Kesehatan Republik Indonesia, model ini dapat diintegrasikan ke dalam platform telemedicine nasional, sehingga meningkatkan akses diagnosis pneumonia di daerah terpencil yang masih kekurangan tenaga radiolog.

1.5 Batasan Masalah

Penelitian ini dibatasi pada Sumber Data yakni Penelitian hanya menggunakan citra *chest X-ray* dari dataset publik Chest X-Ray Images (Pneumonia) di Kaggle (Paul Timothy Mooney). Dataset ini terdiri dari 5.856 citra dalam format JPEG, yang terbagi menjadi dua kelas utama: Normal dan Pneumonia. Seluruh data telah teranotasi oleh ahli radiologi, tanpa penambahan data baru maupun informasi klinis tambahan di luar citra yang tersedia.

BAB II

STUDI PUSTAKA

Pneumonia merupakan penyakit infeksi akut yang menyerang saluran pernapasan bagian bawah dan menyebabkan peradangan pada alveolus, yaitu kantung udara tempat pertukaran oksigen dan karbondioksida berlangsung. Dalam kondisi normal, alveolus memiliki struktur elastis yang memungkinkan difusi gas secara optimal. Namun, ketika terjadi infeksi, ruang ini terisi oleh cairan inflamasi, lendir, atau nanah yang menghambat proses ventilasi paru (WHO, 2020). Pneumonia telah lama menjadi masalah kesehatan global dan termasuk dalam penyebab utama kematian penyakit menular, terutama pada anak-anak di bawah usia lima tahun dan lansia (The Lancet, 2015). Almirall et al. (2017) melaporkan bahwa pneumonia menjadi salah satu penyebab terbanyak pasien dewasa menjalani rawat inap, dengan beban mortalitas yang tetap tinggi meskipun telah tersedia berbagai pilihan pengobatan.

Penyakit ini dapat disebabkan oleh berbagai mikroorganisme patogen, dan karakteristik infeksinya sangat bergantung pada faktor penyebab dan kondisi imunologis individu. Bakteri merupakan etiologi paling umum, terutama *Streptococcus pneumoniae* yang dilaporkan sebagai penyebab utama di banyak negara (Mandell et al., 2019). Patogen lain yang sering ditemukan antara lain *Haemophilus influenzae* dan *Staphylococcus aureus*. Selain bakteri, virus juga menjadi penyebab signifikan, terutama Respiratory Syncytial Virus (RSV), virus Influenza, Adenovirus, dan pada beberapa kasus SARS-CoV-2 seperti pada pandemi COVID-19 (Thorax, 2000; PMC, 2020). Patogen atipikal seperti

Mycoplasma pneumoniae dan *Chlamydothila pneumoniae* juga berperan dalam kasus komunitas dengan gejala ringan hingga sedang. Pada individu dengan imunitas rendah, jamur seperti *Pneumocystis jirovecii* dapat menjadi penyebab infeksi serius.

Proses terjadinya pneumonia umumnya diawali ketika mikroorganisme patogen masuk ke saluran pernapasan melalui droplet yang terhirup, aspirasi sekret orofaring, atau penyebaran hematogen dari infeksi di organ lain (Mandell et al., 2019). Pada individu sehat, sistem pertahanan tubuh seperti silia bronkial, refleks batuk, dan sel imun alveolar berperan mencegah kolonisasi. Namun, jika jumlah patogen berlebihan atau sistem imun melemah, patogen berhasil menembus pertahanan tersebut. Patogenesis berlanjut dengan aktivasi respon imun bawaan, yang memicu pelepasan mediator inflamasi dan migrasi sel-sel imun. Reaksi ini menyebabkan peningkatan permeabilitas kapiler, sehingga cairan plasma dan leukosit masuk ke alveolus dan menimbulkan konsolidasi (The Lancet, 2010). Kondisi ini mengganggu pertukaran gas dan menyebabkan hipoksemia yang ditandai oleh sesak napas dan takipnea.

Faktor risiko memainkan peran penting dalam menentukan kerentanan seseorang terhadap pneumonia. Kelompok usia ekstrem seperti bayi dan balita memiliki sistem imun yang belum matang, sedangkan lansia mengalami penurunan fungsi imun dan elastisitas paru (WHO, 2020). Individu dengan penyakit kronis seperti diabetes, penyakit jantung, gagal ginjal, dan COPD juga memiliki risiko lebih tinggi. Kebiasaan merokok, malnutrisi, konsumsi alkohol, polusi udara, dan kondisi hunian padat memperburuk risiko terjadinya infeksi (The Lancet, 2015).

Selain itu, pasien dengan imunokompromais akibat HIV/AIDS, kemoterapi, atau penggunaan obat imunosupresif lebih rentan mengalami pneumonia berat.

Secara klinis, pneumonia menimbulkan gejala seperti batuk berdahak, demam, menggigil, nyeri dada pleuritik, sesak napas, dan pada anak-anak dapat muncul napas cepat (tachypnea) serta retraksi dinding dada (Hansell et al., 2008). Pada kasus berat, pasien dapat mengalami penurunan saturasi oksigen, hipotensi, atau kebingungan. Komplikasi yang sering muncul meliputi efusi pleura, abses paru, gagal napas akut, hingga sepsis (ScienceDirect, 2010). Pada beberapa kasus, respon inflamasi sistemik yang tidak terkendali dapat menyebabkan kegagalan multi organ yang meningkatkan risiko kematian, terutama pada pasien dengan komorbiditas.

Klasifikasi pneumonia secara klinis sering dibedakan menjadi Community-Acquired Pneumonia (CAP), yaitu infeksi yang terjadi di luar fasilitas kesehatan, dan Hospital-Acquired Pneumonia (HAP), yang muncul setelah lebih dari 48 jam perawatan rumah sakit. Ada juga Ventilator-Associated Pneumonia (VAP), yang terjadi pada pasien yang menggunakan ventilator mekanik (Mandell et al., 2019). Setiap jenis memiliki profil patogen dan prognosis yang berbeda, sehingga pendekatan diagnostik dan terapinya juga harus disesuaikan.

Dalam praktik medis modern, diagnosis pneumonia dilakukan melalui kombinasi pemeriksaan klinis, laboratorium, dan radiologi. Pemeriksaan fisik mencakup auskultasi suara napas yang dapat menunjukkan crackles atau ronki, sedangkan pemeriksaan darah dapat menunjukkan peningkatan leukosit dan penanda inflamasi. Namun, pemeriksaan radiografi dada tetap menjadi alat

diagnosis utama karena mampu menampilkan konsolidasi, infiltrat, atau opasitas pada paru (Hansell et al., 2008). Citra X-ray lebih banyak digunakan karena cepat, murah, dan tersedia luas di fasilitas kesehatan primer maupun sekunder.

Walau demikian, interpretasi radiografi dada tidak selalu mudah karena pentingnya pengalaman dan keterampilan radiolog. Majkowska et al. (2020) menyatakan bahwa variabilitas antar-pemeriksa (interobserver variability) cukup tinggi meskipun dilakukan oleh radiolog ahli. Thorax (2000) juga mencatat bahwa pneumonia dapat menunjukkan pola yang berbeda-beda pada citra, sehingga diagnosis berbasis radiologi kadang sulit dibedakan dari kondisi lain seperti tuberkulosis atau edema paru. Keterlambatan diagnosis atau salah interpretasi dapat meningkatkan risiko komplikasi dan keterlambatan penanganan.

Kondisi tersebut membuka peluang penting bagi implementasi teknologi kecerdasan buatan dalam bidang radiologi. Sistem berbasis deep learning telah dikembangkan untuk mengenali pola pneumonia secara konsisten dari citra X-ray dan memiliki potensi untuk membantu tenaga medis mengambil keputusan yang lebih cepat dan akurat (PMC, 2020). The Lancet (2015) menekankan bahwa integrasi pendekatan digital dalam diagnosis dan pemantauan pneumonia berpotensi mengurangi mortalitas secara signifikan, terutama di negara berkembang yang kekurangan tenaga radiolog.

Dengan beban penyakit yang tinggi, variabilitas diagnosis manual, dan keterbatasan akses layanan kesehatan, pneumonia menjadi salah satu contoh penyakit yang mendesak untuk ditangani menggunakan pendekatan inovatif berbasis teknologi. Kombinasi metode klinis tradisional dan sistem pendukung

keputusan berbasis AI diharapkan dapat meningkatkan kualitas interpretasi citra, mempercepat identifikasi kasus, serta mengurangi risiko kesalahan diagnosis. Oleh karena itu, penelitian terkait otomatisasi deteksi pneumonia melalui citra X-ray berbasis kecerdasan buatan memiliki relevansi klinis dan akademis yang kuat, sekaligus mendukung transformasi layanan kesehatan yang lebih efisien dan akurat.

Convolutional Neural Network (CNN) merupakan arsitektur jaringan saraf tiruan yang dirancang khusus untuk menangani data berbentuk citra. CNN bekerja dengan mengekstraksi fitur visual secara bertahap melalui serangkaian lapisan konvolusi yang mendeteksi pola dasar seperti tepi, tekstur, dan bentuk objek. Lapisan konvolusi ini diikuti oleh *pooling layer* yang berfungsi mereduksi dimensi fitur agar komputasi lebih efisien tanpa menghilangkan informasi penting. Pada tahap akhir, CNN menggunakan *fully connected layer* untuk mengubah hasil ekstraksi fitur menjadi prediksi kelas sehingga model dapat melakukan klasifikasi secara otomatis.

Menurut LeCun et al. (2015) dalam *Nature*, keunggulan utama CNN dibandingkan metode pembelajaran tradisional terletak pada kemampuannya melakukan *feature learning* secara otomatis tanpa perlu rekayasa fitur (*feature engineering*) secara manual. Dalam konteks citra medis seperti X-ray paru-paru, CNN mampu mengenali perbedaan visual halus antara jaringan paru normal dan area yang mengalami infiltrasi akibat pneumonia.

Dalam implementasi praktis, teknik *transfer learning* banyak digunakan untuk mempercepat pelatihan CNN dan meningkatkan akurasi, terutama pada dataset medis yang umumnya terbatas. Transfer learning dilakukan dengan

memanfaatkan model pra-latih seperti EfficientNet, MobileNet, ResNet, atau DenseNet yang sebelumnya telah dilatih pada dataset berskala besar seperti ImageNet. Model tersebut kemudian disesuaikan kembali pada tugas klasifikasi pneumonia dengan hanya mengganti lapisan akhir sesuai jumlah kelas target. Studi oleh Kermayn et al. (2018) menunjukkan bahwa penggunaan transfer learning pada model InceptionV3 berhasil mencapai akurasi deteksi pneumonia di atas 90%, bahkan dengan jumlah data pelatihan yang terbatas. Dengan demikian, CNN merupakan fondasi utama dalam pengembangan sistem deteksi pneumonia berbasis citra X-ray, dan transfer learning menjadi strategi penting untuk meningkatkan performanya secara efisien.

Ensemble learning merupakan pendekatan dalam pembelajaran mesin yang bertujuan meningkatkan performa model dengan menggabungkan beberapa algoritma secara simultan. Berbeda dengan model tunggal seperti CNN yang hanya mengandalkan satu arsitektur dalam melakukan prediksi, ensemble learning memanfaatkan keragaman dari beberapa model untuk menghasilkan keputusan yang lebih stabil dan akurat. Prinsip dasarnya mirip dengan pengambilan keputusan kolektif, di mana kesalahan satu model dapat dikompensasi oleh model lainnya, sehingga hasil akhir menjadi lebih andal (Zhou, 2012).

Terdapat tiga pendekatan utama dalam ensemble learning, yaitu bagging, boosting, dan stacking. Bagging (*bootstrap aggregating*) bekerja dengan melatih beberapa model secara paralel menggunakan subset data yang diambil secara acak dengan pengembalian. Metode ini efektif untuk mengurangi variansi dan mencegah overfitting, seperti pada algoritma Random Forest (Breiman, 2001). Sebaliknya,

boosting melatih model secara berurutan, di mana setiap model fokus memperbaiki kesalahan model sebelumnya. Teknik ini mampu meningkatkan akurasi secara signifikan, namun cenderung lebih rentan terhadap noise dan membutuhkan waktu pelatihan lebih lama (Schapire, 1999).

Stacking ensemble, yang digunakan dalam penelitian ini, bekerja dengan menggabungkan keluaran dari beberapa model dasar (*base learners*) dan memprosesnya kembali menggunakan model tambahan yang disebut *meta-learner*. Pendekatan ini tidak sekadar melakukan voting atau averaging, tetapi belajar secara eksplisit bagaimana mengombinasikan prediksi model dasar secara optimal. Menurut Wolpert (1992), stacking memiliki fleksibilitas yang lebih tinggi dibandingkan bagging dan boosting karena memungkinkan penggabungan model dengan arsitektur berbeda. Dalam konteks deteksi pneumonia berbasis citra X-ray, stacking telah terbukti memberikan peningkatan akurasi dibandingkan model tunggal, sebagaimana ditunjukkan dalam penelitian Liang dan Yu (2020) serta Hashmi et al. (2020) yang melaporkan nilai akurasi di atas 92%.

Dengan demikian, ensemble learning khususnya stacking menjadi solusi yang tepat untuk meningkatkan kinerja model klasifikasi pneumonia, karena mampu memanfaatkan kelebihan masing-masing arsitektur CNN sambil meminimalkan kelemahan individu model.

Penelitian oleh Rajpurkar et al. (2017) mengembangkan model CNN berbasis arsitektur CheXNet untuk mendeteksi pneumonia pada citra X-ray paru-paru dari dataset ChestX-ray14. CheXNet, yang merupakan modifikasi dari DenseNet-121, dilatih untuk mengklasifikasikan 14 jenis kelainan toraks, termasuk

pneumonia. Hasil penelitian menunjukkan bahwa CheXNet mencapai performa setara atau lebih baik dibandingkan radiolog dengan nilai AUC-ROC sebesar 0.76 untuk deteksi pneumonia. Namun, model ini menghadapi tantangan dalam menangani variasi kualitas citra dan ketidakseimbangan kelas, sehingga memerlukan teknik augmentasi data untuk meningkatkan generalisasi model. Kelemahan lainnya adalah kompleksitas komputasi yang tinggi, yang dapat menyulitkan penerapan pada perangkat dengan sumber daya terbatas.

Liang dan Yu (2020) mengusulkan pendekatan stacking ensemble untuk deteksi pneumonia pada citra X-ray paru-paru menggunakan dataset publik dari Kaggle (Chest X-Ray Pneumonia). Peneliti menggabungkan tiga model dasar, yaitu VGG16, ResNet50, dan MobileNetV2, dengan meta-learner berbasis logistic regression. Pendekatan ini memanfaatkan kelebihan masing-masing model untuk menangani variasi pola pneumonia seperti infiltrasi dan konsolidasi. Hasil pengujian menunjukkan peningkatan performa dibandingkan model tunggal, dengan nilai akurasi 92%, sensitivitas 90%, spesifisitas 93%, F1-score 91%, dan AUC-ROC 0.95. Augmentasi data seperti rotasi, flipping, dan penyesuaian kontras diterapkan untuk meningkatkan ketahanan model terhadap variasi citra. Namun, waktu pelatihan meningkat signifikan karena kompleksitas penggabungan model, yang menjadi tantangan untuk aplikasi real-time.

Penelitian oleh Kermany et al. (2018) menggunakan pendekatan transfer learning dengan model InceptionV3 untuk mendeteksi pneumonia pada citra X-ray anak-anak. Dataset yang digunakan berasal dari Guangzhou Women and Children's Medical Center, yang terdiri dari kelas normal dan pneumonia. Model dilatih

dengan teknik augmentasi data seperti random cropping dan rotasi untuk mengatasi keterbatasan jumlah data. Hasilnya menunjukkan akurasi 92.8%, sensitivitas 93.2%, spesifisitas 90.1%, dan AUC-ROC 0.94. Meskipun performanya menjanjikan, model ini kurang optimal dalam menangani citra dengan kualitas rendah atau noise, sehingga peneliti menyarankan pengembangan metode ensemble untuk meningkatkan ketahanan model.

Sari et al. (2021) mengembangkan model berbasis EfficientNetB0 untuk deteksi pneumonia pada citra X-ray paru-paru dengan dataset Chest X-Ray Pneumonia dari Kaggle. Model ini dipilih karena efisiensinya dalam mengurangi jumlah parameter dibandingkan arsitektur lain seperti ResNet atau VGG. Hasil pengujian menunjukkan akurasi 90%, sensitivitas 88%, spesifisitas 91%, F1-score 89%, dan AUC-ROC 0.93. Peneliti juga menerapkan augmentasi data seperti scaling dan shearing untuk meningkatkan generalisasi model. Namun, model tunggal ini masih menghadapi tantangan dalam mendeteksi kasus pneumonia dengan pola yang tidak jelas, sehingga peneliti menyarankan penggunaan pendekatan ensemble untuk meningkatkan performa.

Menurut Chouhan et al. (2020), pendekatan ensemble learning dapat meningkatkan akurasi deteksi pneumonia dengan menggabungkan beberapa model CNN seperti AlexNet, DenseNet, dan ResNet18. Penelitian ini menggunakan dataset publik dan menerapkan stacking ensemble dengan meta-learner berbasis Support Vector Machine (SVM). Hasilnya menunjukkan akurasi 93.5%, sensitivitas 92%, spesifisitas 94%, F1-score 92.8%, dan AUC-ROC 0.96. Keunggulan pendekatan ini adalah kemampuan menangani variasi data yang

kompleks, tetapi kelemahannya adalah peningkatan kompleksitas komputasi dan waktu pelatihan. Peneliti menyarankan optimasi hyperparameter meta-learner untuk meningkatkan efisiensi.

Penelitian oleh Ayan dan Ünver (2019) mengeksplorasi kombinasi MobileNetV2 dan Xception dalam pendekatan stacking ensemble untuk deteksi pneumonia. Dataset yang digunakan adalah Chest X-Ray Pneumonia dari Kaggle, yang dibagi menjadi 70% untuk pelatihan, 20% untuk validasi, dan 10% untuk pengujian. Augmentasi data seperti rotasi dan zooming diterapkan untuk meningkatkan ketahanan model. Meta-learner berbasis jaringan saraf sederhana digunakan untuk menggabungkan output model dasar. Hasilnya menunjukkan akurasi 91%, sensitivitas 89%, spesifisitas 92%, F1-score 90%, dan AUC-ROC 0.94. Kelemahan utama adalah kebutuhan akan sumber daya komputasi yang lebih besar dibandingkan model tunggal, sehingga kurang ideal untuk perangkat dengan spesifikasi rendah.

Dalam penelitian oleh Hashmi et al. (2020), pendekatan stacking ensemble digunakan dengan menggabungkan model EfficientNetB0, MobileNetV3Small, dan InceptionV3, dengan meta-learner berbasis Random Forest. Dataset Chest X-Ray Pneumonia dari Kaggle digunakan, dengan praproses data meliputi normalisasi intensitas piksel dan augmentasi seperti flipping dan rotasi. Hasil pengujian menunjukkan akurasi 94%, sensitivitas 93%, spesifisitas 95%, F1-score 93.5%, dan AUC-ROC 0.97. Pendekatan ini menunjukkan performa yang lebih baik dibandingkan model tunggal, tetapi kompleksitas model meningkatkan waktu pelatihan, yang menjadi tantangan untuk aplikasi skala besar.

Berdasarkan tinjauan di atas, metrik evaluasi seperti akurasi, sensitivitas, spesifisitas, F1-score, dan AUC-ROC menjadi standar untuk mengukur performa model deteksi pneumonia. Penelitian-penelitian sebelumnya menunjukkan bahwa pendekatan stacking ensemble cenderung menghasilkan performa yang lebih baik dibandingkan model tunggal karena kemampuannya menggabungkan kekuatan berbagai arsitektur CNN. Namun, tantangan seperti peningkatan kompleksitas komputasi dan waktu pelatihan masih menjadi kendala, terutama untuk aplikasi real-time atau pada perangkat dengan sumber daya terbatas. Penelitian ini akan menggunakan model dasar EfficientNetB0, MobileNetV3Small, dan MobileNetV2 dengan meta-learner berbasis jaringan saraf sederhana, untuk mengatasi tantangan tersebut dan meningkatkan akurasi deteksi pneumonia pada citra X-ray paru-paru.

Rangkuman hasil evaluasi performa model dari penelitian sebelumnya disajikan pada Tabel 2.1 berikut:

Tabel 2. 1 Rangkuman Performa Model Deteksi Pneumonia

Peneliti	Arsitektur	Datas et (Sum ber)	Juml ah Data (Total)	Aku rasi (%)	Se nsi tivi tas (%)	Spe sifi sitas (%)	F1- Score (%)	AUC- ROC
Rajpurkar et al. (2017)	CheXNet (DenseNet-121)	Chest X-ray14 (NIH)	112.120 citra	-	-	-	-	0.76
Liang & Yu (2020)	Stacking (VGG16, ResNet50, MobileNetV2)	Chest X-Ray Pneumonia (Kaggle)	5.856 citra	92	90	93	91	0.95

Peneliti	Arsitektur	Datas et (Sum ber)	Juml ah Data (Total)	Aku rasi (%)	Se nsi tivi tas (%)	Spe sifis itas (%)	F1- Score (%)	AUC- ROC
Kerman y et al. (2018)	InceptionV3	Guan gzhou Pediat rics Datas et	5.863 citra	92.8	93. 2	90.1	-	0.94
Sari et al. (2021)	EfficientNet B0	Chest X- Ray Pneu monia (Kagg le)	± 5.80 0 citra	90	88	91	89	0.93
Chouha n et al. (2020)	Stacking (AlexNet, DenseNet, ResNet18)	Chest X- Ray Pneu monia (Kagg le)	± 5.80 0 citra	93.5	92	94	92.8	0.96
Ayan & Ünver (2019)	Stacking (MobileNet V2, Xception)	Chest X- Ray Pneu monia (Kagg le)	5.856 citra	91	89	92	90	0.94
Hashmi et al. (2020)	Stacking (EfficientNe tB0, MobileNetV 3Small, InceptionV3)	Chest X- Ray Pneu monia (Kagg le)	5.856 citra	94	93	95	93.5	0.97

Berdasarkan tinjauan penelitian terdahulu, pendekatan CNN tunggal seperti InceptionV3 atau EfficientNet terbukti mampu mendeteksi pneumonia dengan

akurasi tinggi. Namun, beberapa penelitian seperti yang dilaporkan oleh Lakhani dan Sundaram (2017) serta Zech et al. (2018) menunjukkan bahwa performa model tunggal cenderung menurun ketika diuji pada dataset atau populasi yang berbeda, sehingga memiliki keterbatasan dalam hal generalisasi. Sebaliknya, penelitian dengan pendekatan stacking ensemble seperti yang dilakukan oleh Liang dan Yu (2020), Chouhan et al. (2020), dan Hashmi et al. (2020) secara konsisten melaporkan peningkatan akurasi dan sensitivitas dibandingkan model tunggal. Hal ini menunjukkan bahwa penggabungan beberapa model dasar mampu menangkap variasi pola citra pneumonia secara lebih komprehensif.

Namun demikian, sebagian besar penelitian tersebut hanya membuktikan keunggulan stacking secara empiris tanpa menjelaskan secara konseptual mengapa stacking lebih tepat digunakan dibandingkan teknik ensemble lain seperti bagging atau boosting. Oleh karena itu, penelitian ini tidak hanya mengadopsi stacking ensemble sebagai pendekatan utama, tetapi juga menegaskan landasan metodologis pemilihannya sebagai strategi yang lebih adaptif dalam menggabungkan kekuatan berbagai arsitektur CNN untuk deteksi pneumonia berbasis citra X-ray.

BAB III

DESAIN DAN IMPLEMENTASI SISTEM KLASIFIKASI PNEUMONIA

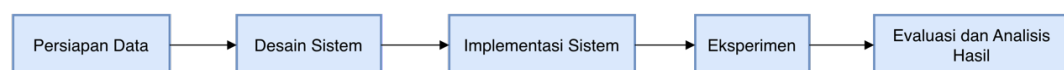
3.1 Prosedur Penelitian

Penelitian ini menggunakan metode eksperimen kuantitatif dengan pendekatan *ensemble learning*, khususnya teknik *stacking ensemble*, untuk mendeteksi pneumonia dari citra X-ray paru-paru. Desain penelitian ini mencakup beberapa tahapan utama, yaitu: pengumpulan data, desain system *stacking ensemble*, eksperimen dan evaluasi model, serta evaluasi performa sistem. Dataset yang digunakan adalah *Chest X-Ray Pneumonia* dari Kaggle, yang terdiri dari dua kelas, yaitu *Normal* dan *Pneumonia*.

Backbone yang digunakan dalam penelitian ini adalah EfficientNetB0, MobileNetV2, dan MobileNetV3Small. Ketiga arsitektur ini dipilih karena mampu memberikan keseimbangan antara akurasi dan efisiensi komputasi, sehingga relevan untuk implementasi di lingkungan klinis dengan keterbatasan sumber daya. EfficientNetB0 memiliki pendekatan *compound scaling* yang mengoptimalkan kedalaman, lebar, dan resolusi jaringan sehingga dapat mencapai akurasi tinggi dengan jumlah parameter relatif kecil (Tan & Le, 2019). MobileNetV2 menggunakan blok *inverted residual* dan *linear bottleneck* yang membuatnya efisien dalam mengekstraksi fitur citra dengan ukuran model yang ringan (Sandler et al., 2018). Sementara itu, MobileNetV3Small dirancang melalui *Neural Architecture Search* (NAS) dan dilengkapi modul *Squeeze-and-Excitation* serta aktivasi *hard-swish* yang lebih efisien secara komputasi, menjadikannya sangat cocok untuk perangkat *edge computing* (Howard et al., 2019). Walaupun

MobileNetV2 dan MobileNetV3Small berasal dari rumpun arsitektur yang sama, keduanya tetap memberikan keragaman representasi fitur berkat perbedaan desain blok dasar dan fungsi aktivasi. Dengan kombinasi ini, meta-learner pada tahap stacking ensemble dapat memanfaatkan komplementaritas antar backbone, sehingga meningkatkan generalisasi model dalam mendeteksi pneumonia.

Masing-masing model dilatih secara terpisah dan digunakan untuk menghasilkan prediksi probabilitas terhadap kelas target. Probabilitas dari ketiga model kemudian digabungkan dan digunakan sebagai input ke dalam model meta-learner berupa jaringan saraf sederhana. Meta-learner ini dilatih untuk mempelajari kombinasi optimal dari output model dasar. Evaluasi dilakukan menggunakan metrik akurasi, sensitivitas, dan spesifisitas untuk mengukur performa baik dari model individual maupun sistem ensemble. Diagram alir metodologi penelitian disajikan pada Gambar 3.1.



Gambar 3. 1 Desain Penelitian

3.2 Persiapan Data

Dataset yang digunakan dalam penelitian ini adalah *Chest X-Ray Images (Pneumonia)* yang tersedia pada platform Kaggle. Dataset ini berisi citra X-ray dada dalam format JPEG dengan total 5.863 citra, yang terbagi ke dalam dua kelas, yaitu NORMAL (label 0) sebanyak 1.583 citra, dan PNEUMONIA (label 1) sebanyak 4.273 citra. Dengan demikian, komposisi dataset bersifat imbalanced, di mana kelas PNEUMONIA mendominasi sekitar 73%, sedangkan kelas NORMAL hanya mencakup 27% dari keseluruhan data.

Dataset ini telah dibagi secara default oleh penyedia menjadi tiga subset, yaitu training set, validation set, dan test set dengan distribusi sebagai berikut:

Tabel 3. 1 Sebaran dataset

Subset	Total Citra	NORMAL (0)	PNEUMONIA (1)	Proporsi
Training Set	5.216	1.341	3.875	26% : 74%
Validation Set	16	8	8	50% : 50%
Test Set	624	234	390	38% : 62%

Distribusi tersebut menunjukkan bahwa terdapat ketidakseimbangan kelas (*class imbalance*) terutama pada bagian training set. Jika dibiarkan, model cenderung bias terhadap kelas mayoritas (PNEUMONIA) sehingga berpotensi menghasilkan sensitivitas tinggi tetapi spesifisitas rendah.

Untuk mengatasi hal tersebut, penelitian ini tidak melakukan augmentasi visual agar tetap menjaga keaslian citra. Sebagai gantinya, *class weighting* diterapkan pada tahap pelatihan model dengan memberikan bobot lebih besar pada kelas minoritas (NORMAL/0). Bobot kelas dihitung menggunakan rumus:

$$\text{class_weight}(i) = \frac{K \times n_i}{N} \quad (3.1)$$

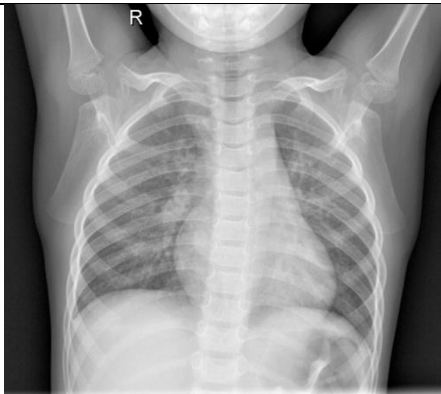
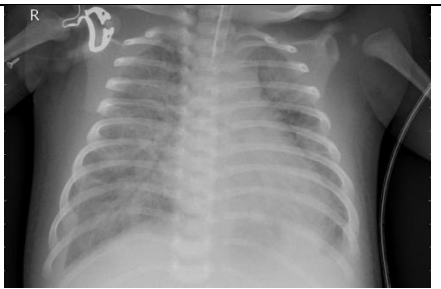
dengan N adalah jumlah total data, K adalah jumlah kelas, dan n_i adalah jumlah sampel pada kelas ke- i . Berdasarkan perhitungan, diperoleh bobot:

- Weight_NORMAL (0) ≈ 1.85

- $\text{Weight_PNEUMONIA (1)} \approx 0.69$

Dengan penerapan class weighting ini, model tetap belajar secara proporsional tanpa perlu melakukan penambahan citra secara sintetis. Strategi class weighting dipilih karena banyak digunakan dalam penelitian medis sebagai pendekatan untuk menangani *imbalanced data* tanpa mengubah struktur asli citra diagnostik. Pendekatan ini terbukti efektif pada beberapa studi radiologi seperti Johnson et al. (2019), Cohen et al. (2020), dan Ribli et al. (2018) yang lebih memilih *loss rebalancing* dibandingkan augmentasi visual yang berpotensi mengganggu karakteristik klinis citra. Berikut adalah contoh gambar dari label Normal dan Pneuomonia ditunjukkan pada Tabel 3.2

Tabel 3. 2 Contoh gambar dataset citra X-ray

No	Gambar	Label
1.		Normal
2.		Pneuomonia

3.3 Desain Sistem

Desain sistem dibuat untuk menggambarkan secara umum bagaimana sistem melakukan klasifikasi Pneumonia menggunakan pendekatan Stacking Ensemble. Desain sistem untuk penelitian ini ditunjukkan pada Gambar 3.2.



Gambar 3. 2 Desain Sistem

3.3.1 Input Image

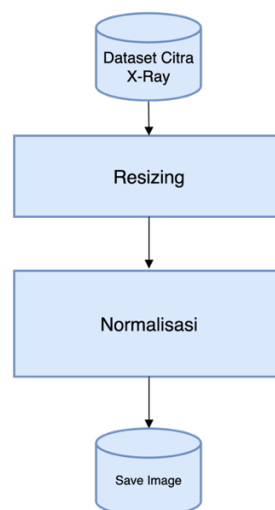
Citra radiografi toraks menjadi masukan utama pada sistem deteksi pneumonia yang dikembangkan. Data citra diperoleh dari dataset Chest X-ray Images (Pneumonia) yang tersedia di repositori Kaggle. Dataset ini secara khusus dirancang untuk deteksi penyakit paru-paru dan telah banyak digunakan dalam penelitian sebelumnya. Dataset terdiri dari 5.863 citra X-ray dada dalam format JPEG, yang masing-masing telah dilabeli ke dalam dua kelas utama: Pneumonia, untuk citra pasien yang terdiagnosis pneumonia secara klinis, dan Normal, untuk citra pasien sehat atau tanpa indikasi pneumonia.

Seluruh citra pada dataset ini merupakan hasil radiografi toraks dalam posisi *posteroanterior (PA)*, dan label *ground truth* berasal dari interpretasi radiologis yang telah divalidasi oleh ahli radiologi.

Sebelum digunakan pada model, citra-citra ini melalui tahap praproses guna memastikan keseragaman ukuran, format, dan kualitas agar kompatibel dengan arsitektur *Convolutional Neural Network (CNN)* yang digunakan. Penjelasan rinci mengenai tahapan praproses, seperti pengubahan ukuran citra dan konversi tipe data, disajikan pada subbab 3.3.2 Preprocessing Citra.

3.3.2 Preprocessing Citra

Pada tahap ini citra masukan diubah ukurannya menjadi resolusi seragam (224x224 piksel), dikonversi ke format warna RGB, dan direpresentasikan sebagai array numerik bertipe float32, langkah praproses lanjutan dilakukan untuk mengoptimalkan data bagi model pembelajaran mendalam. Praproses ini esensial untuk meningkatkan kualitas data, memperluas variasi, dan menyesuaikannya dengan persyaratan masukan arsitektur jaringan saraf. Ada tiga tahap dalam preprocessing citra pada penelitian ini yang disajikan pada Gambar 3.4.



Gambar 3. 3 *Preprocessing* Citra

3.3.2.1 Resizing

Resizing merupakan proses pengubahan ukuran citra dari resolusi asli menjadi ukuran standar yang telah ditentukan. Dalam penelitian ini, seluruh citra X-ray diubah ukurannya menjadi resolusi **224×224 piksel**. Pemilihan ukuran ini didasarkan pada persyaratan arsitektur *Convolutional Neural Network (CNN)* yang digunakan, di mana model *backbone* seperti EfficientNetB0, MobileNetV3Small,

dan MobileNetV2 memerlukan masukan dengan dimensi yang seragam.

Proses resizing dilakukan dengan metode interpolasi piksel, yaitu memperkirakan nilai piksel baru berdasarkan nilai piksel-piksel tetangganya. Jika ukuran citra diperbesar, interpolasi mengisi nilai piksel tambahan di antara piksel-piksel asli. Sebaliknya, jika ukuran citra diperkecil, interpolasi memilih dan menggabungkan nilai dari beberapa piksel asli untuk menghasilkan piksel baru.

Dengan melakukan resizing, setiap citra memiliki dimensi masukan yang konsisten sehingga memudahkan proses pelatihan, mempercepat komputasi, dan mencegah kesalahan pemrosesan akibat perbedaan ukuran citra. Selain itu, penyamaan ukuran citra juga membantu model fokus pada pola visual yang relevan tanpa dipengaruhi oleh variasi skala pada data asli.

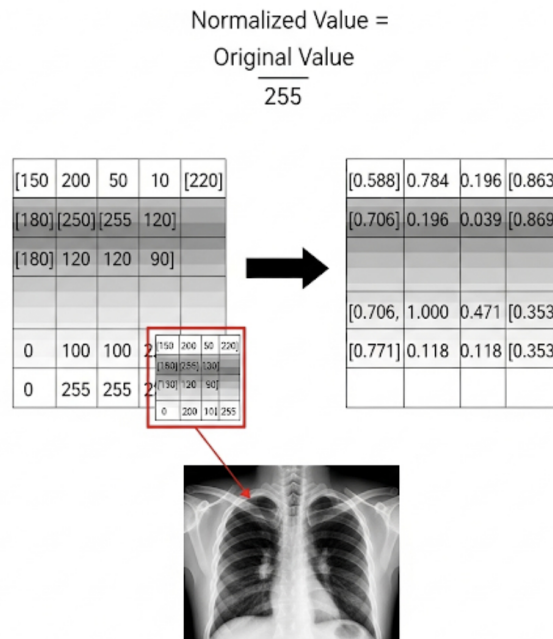
3.3.2.2 Normalization

Setelah citra dikonversi menjadi array bertipe **float32**, nilai piksel yang semula berada pada rentang 0–255 diubah ke skala yang lebih terstandarisasi.

Tujuan normalisasi meliputi:

1. Menyamakan skala fitur agar setiap piksel memiliki kontribusi yang sebanding dalam proses pelatihan, sehingga model tidak bias terhadap nilai piksel yang lebih besar.
2. Mempercepat konvergensi pelatihan, data yang terstandarisasi cenderung membuat optimisasi lebih stabil dan memerlukan *epoch* yang lebih sedikit.
3. Menyesuaikan dengan bobot *pre-trained model*, model *backbone* seperti EfficientNetB0, MobileNetV3Small, dan MobileNetV2 dilatih pada dataset

ImageNet yang memiliki distribusi nilai tertentu, sehingga data masukan perlu disesuaikan agar kompatibel.



Gambar 3. 4 Normalisasi Nilai Piksel

Secara matematis, proses normalisasi sederhana dapat dinyatakan sebagai:

$$I' = \frac{I}{255} \quad (3.2)$$

dengan I adalah nilai piksel asli (0–255) dan I' adalah nilai piksel hasil normalisasi (0–1).

Pada beberapa model *pre-trained*, normalisasi dilakukan dengan *mean subtraction* dan *standard deviation scaling*:

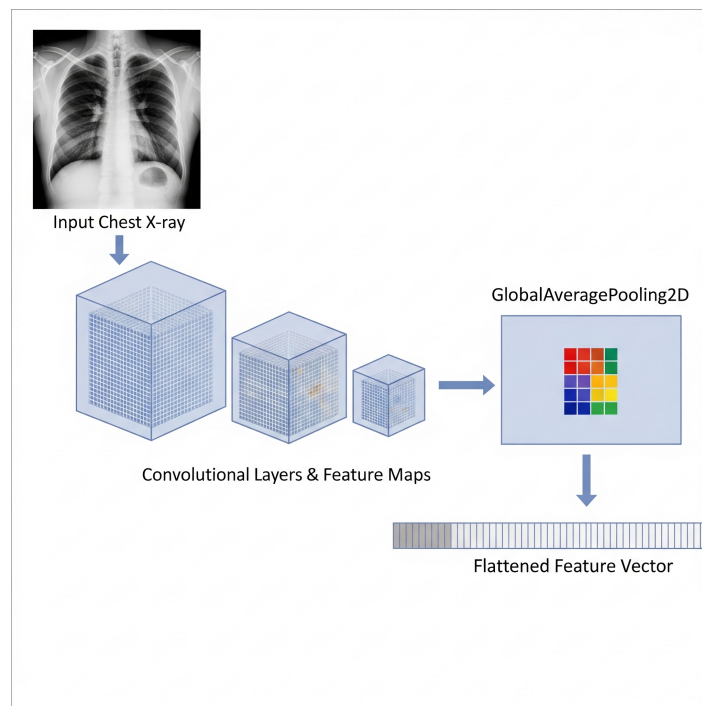
$$I' = \frac{I - \mu}{\sigma} \quad (3.3)$$

di mana μ adalah rata-rata nilai piksel dan σ adalah simpangan baku nilai piksel pada dataset pelatihan asli model. Dengan cara ini, distribusi piksel citra baru akan serupa dengan distribusi data yang digunakan saat pelatihan awal model. Tahap normalisasi memastikan bahwa citra masukan memiliki distribusi nilai yang konsisten dengan parameter pelatihan awal, sehingga model dapat memanfaatkan bobot *pre-trained* secara optimal dan menghasilkan performa yang lebih baik.

3.3.3 Ekstraksi Fitur

Ekstraksi fitur merupakan tahapan yang bertujuan mengubah citra X-ray hasil praproses menjadi representasi numerik dengan dimensi yang lebih ringkas namun sarat informasi. Representasi ini digunakan sebagai dasar bagi proses klasifikasi pada tahap selanjutnya. Pada penelitian ini, proses ekstraksi dilakukan menggunakan tiga arsitektur *backbone* pre-trained, yaitu EfficientNetB0, MobileNetV3Small, dan MobileNetV2. Ketiga backbone ini telah dilatih sebelumnya pada dataset berskala besar (ImageNet) sehingga mampu mengenali beragam pola visual, meskipun citra yang digunakan pada penelitian ini berasal dari domain medis.

Citra X-ray hasil praproses beresolusi 224×224 piksel, format RGB, dan bertipe data float32 diberikan sebagai masukan ke *backbone*. ini memastikan bahwa lapisan *fully connected* bawaan ImageNet dihapus, sehingga keluaran *backbone* bukan berupa prediksi kelas, melainkan peta fitur (*feature maps*) dua dimensi. *Gambar 3.5* menyajikan ilustrasi konseptual proses ekstraksi fitur menggunakan backbone pre-trained



Gambar 3. 5 Ekstraksi Fitur *Pre Trained Model*

Pada tahap awal, lapisan konvolusi pertama dari backbone mendeteksi fitur-fitur dasar seperti tepi, garis, dan perbedaan intensitas. Seiring bertambahnya kedalaman lapisan, jaringan saraf mengekstraksi fitur yang lebih kompleks, termasuk pola-pola radiografis yang khas pada paru-paru seperti infiltrat, konsolidasi, atau tekstur jaringan normal. Proses ini dilakukan secara hierarkis: lapisan awal fokus pada detail lokal, sedangkan lapisan yang lebih dalam menangkap konteks global.

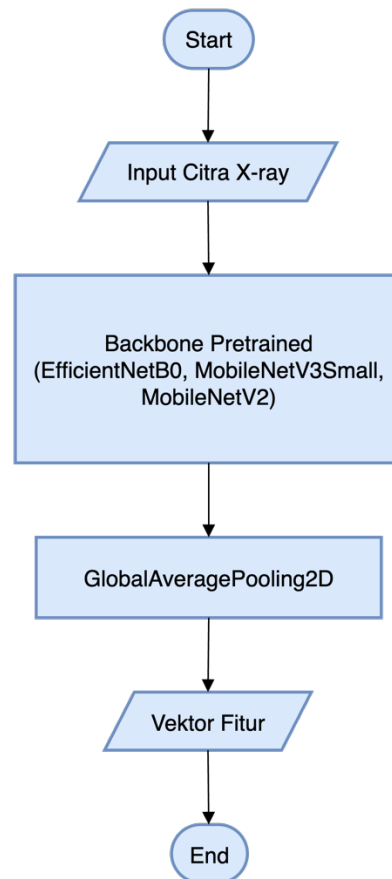
Keluaran akhir dari lapisan konvolusi berupa peta fitur dengan dimensi spasial tertentu ($\text{tinggi} \times \text{lebar} \times \text{jumlah kanal}$). Peta fitur ini kemudian diproses menggunakan lapisan *GlobalAveragePooling2D*, yang menghitung nilai rata-rata setiap kanal di seluruh posisi pikselnya. Hasilnya adalah vektor satu dimensi yang panjangnya setara dengan jumlah kanal pada peta fitur akhir. Panjang vektor ini

berbeda untuk setiap backbone: 1280 elemen pada EfficientNetB0 dan MobileNetV2, serta 576 elemen pada MobileNetV3Small.

Penggunaan GlobalAveragePooling2D memberikan dua keuntungan utama:

1. Reduksi dimensi Mengubah peta fitur spasial menjadi vektor 1D sehingga jumlah parameter yang harus diproses pada tahap klasifikasi menjadi lebih sedikit.
2. Invarian posisi Menghilangkan ketergantungan pada lokasi spesifik suatu fitur dalam citra, sehingga model fokus pada ada atau tidaknya fitur tersebut secara keseluruhan.

Tahap ekstraksi fitur ini menghasilkan vektor fitur terkompresi yang berisi informasi penting dari citra X-ray. Vektor ini kemudian diteruskan ke tahap klasifikasi yang akan dibahas pada sub-bab selanjutnya. *Gambar 3.6* menampilkan diagram alir (flowchart) yang menjelaskan alur tahapan dari input citra hingga dihasilkan vektor fitur.

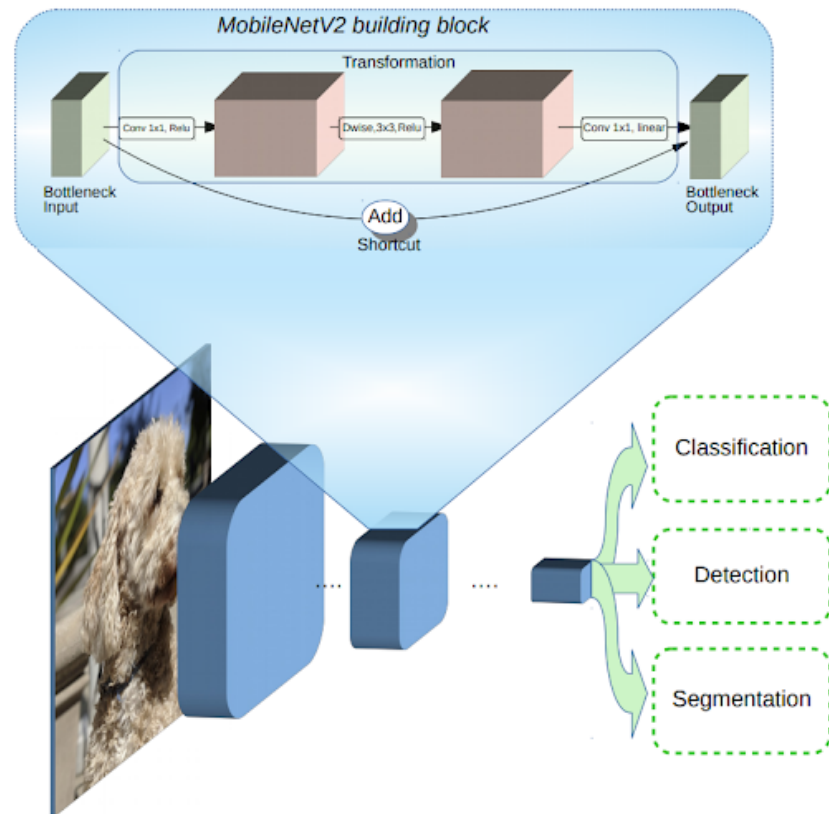


Gambar 3. 6 Flowchart ekstraksi fitur

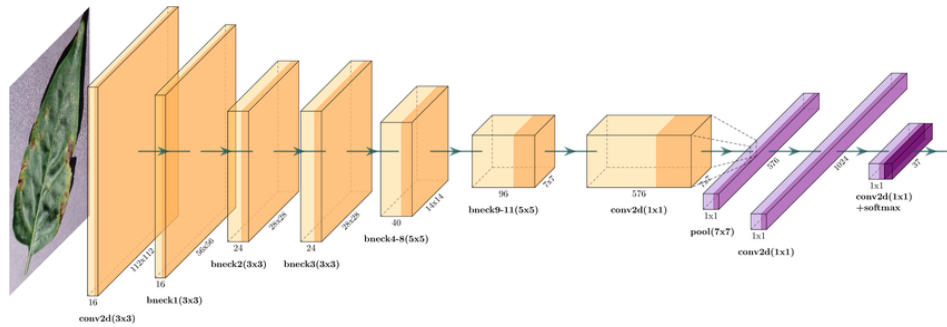
3.3.4 Classifier

Classifier dalam penelitian ini tidak hanya berfungsi sebagai modul matematis untuk memetakan fitur menjadi label, tetapi juga sebagai tahap interpretatif yang menerjemahkan representasi radiologis menjadi keputusan klinis. Sebelum memasuki classifier, citra X-ray diproses oleh tiga backbone berbeda, yaitu EfficientNetB0, MobileNetV2, dan MobileNetV3Small. Ketiganya sengaja dipilih karena meskipun sama-sama Convolutional Neural Network, masing-masing memiliki kecenderungan fokus ekstraksi fitur yang berbeda: EfficientNetB0 menangkap keseimbangan pola global dan lokal melalui mekanisme compound

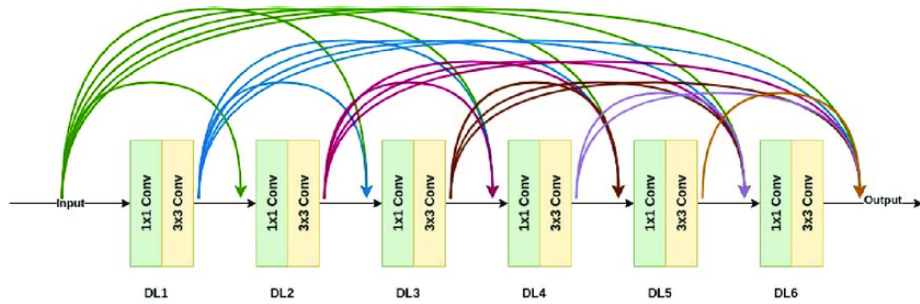
scaling, MobileNetV2 lebih sensitif terhadap tekstur halus seperti infiltrasi bercak-bercak, sedangkan MobileNetV3Small dengan mekanisme channel attention melalui squeeze-and-excitation cenderung menyoroti opasitas dominan khas konsolidasi pneumonia. Dengan demikian, sistem memperoleh tiga perspektif berbeda atas citra X-ray, layaknya tiga radiolog yang melihat dari sudut yang berbeda.



Gambar 3. 7 Arsitektur MobileNetV2 (Sandler et al., 2018)



Gambar 3. 8 Arsitektur MobileNetV3-Small (Howard et al., 2019)



Gambar 3. 9 Arsitektur EfficientNet-B0 (Tan & Le, 2019)

Feature map keluaran dari masing-masing backbone kemudian diringkas menggunakan Global Average Pooling, sebagaimana ditunjukkan pada Persamaan (x):

$$g_k = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W F_k(i, j) \quad (3.4)$$

Proses ini menyaring informasi radiologis penting dari seluruh area paru tanpa bergantung pada lokasi pasti lesi, sehingga cocok digunakan untuk mendeteksi pneumonia yang polanya bisa tersebar maupun terlokalisasi. Vektor fitur yang dihasilkan kemudian diteruskan ke lapisan Dense dengan aktivasi ReLU yang bertugas mempelajari hubungan non-linear antar fitur, misalnya membedakan

apakah infiltrasi ringan pada kedua paru lebih berpotensi pneumonia viral dibanding konsolidasi kuat di satu sisi. Stabilitas pelatihan dijaga dengan Batch Normalization, sedangkan Dropout digunakan untuk mengurangi risiko overfitting.

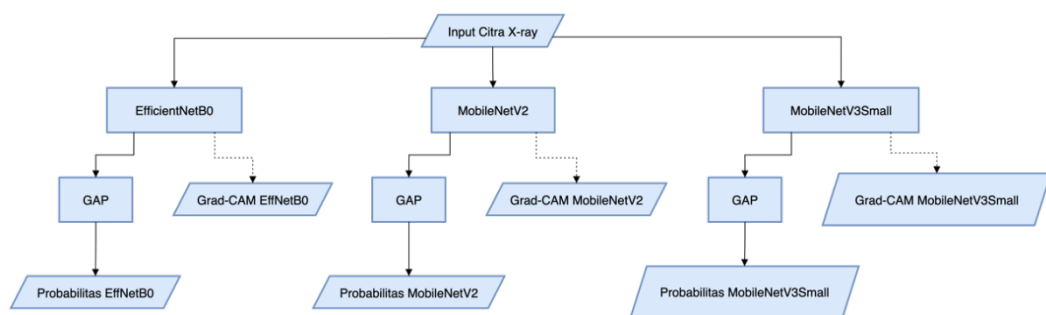
Keputusan akhir dihasilkan melalui lapisan Dense dengan fungsi aktivasi Softmax:

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^K e^{z_k}}, \quad K = 2 \quad (3.5)$$

yang menghasilkan probabilitas terukur terhadap dua kelas target, yaitu Normal dan Pneumonia. Output probabilitas ini dapat diinterpretasikan sebagai tingkat keyakinan klinis model misalnya nilai 0.92 pada kelas Pneumonia menunjukkan keyakinan tinggi setara opini radiolog. Untuk memastikan bahwa sistem tidak sepenuhnya bersifat *black box*, Pendekatan ini memastikan bahwa *classifier* tidak hanya akurat secara numerik, tetapi juga valid secara radiologis dan dapat diterima dalam konteks medis nyata.

Pada penelitian ini juga, output probabilitas dari lapisan Softmax pada masing-masing *backbone* tidak langsung digunakan sebagai prediksi akhir, melainkan dijadikan sebagai *meta-features*. Setiap model dasar menghasilkan dua nilai probabilitas, yaitu P(Normal) dan P(Pneumonia), sehingga tiga *backbone* menghasilkan total enam fitur numerik. Keenam nilai ini kemudian dijadikan input bagi *meta-learner* berupa *Neural Network (NN)* sederhana yang terdiri dari lapisan Dense bertingkat. Meta-learner bertugas mempelajari pola kombinasi antar *backbone*, misalnya mengenali kondisi ketika dua model sangat yakin pada label pneumonia sementara satu model ragu-ragu. Dengan pendekatan ini, proses

ensemble bukan sekadar *voting*, tetapi pengambilan keputusan terarah berbasis pembelajaran kembali atas distribusi probabilitas antar-model. Dengan demikian, classifier pada tiap backbone berperan sebagai *pemberi opini klinis awal*, sementara meta-learner bertindak sebagai *konsultan final* yang menyusun keputusan kolektif berdasarkan keyakinan relatif dari ketiga model.



Gambar 3. 10 Classifier (Base Model)

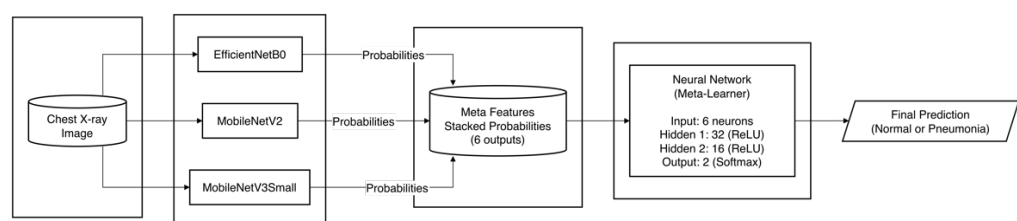
3.3.5 Stacking Ensemble

Tahap Ensemble Stacking adalah inti dari pendekatan ini untuk mencapai performa klasifikasi yang lebih unggul dibandingkan dengan model tunggal. Proses ini tidak hanya mengacu pada kombinasi sederhana dari prediksi, melainkan melibatkan arsitektur dua tingkat di mana prediksi dari model base digunakan sebagai input untuk model tingkat kedua, yang disebut meta-learner. Tahap ini dapat dipecah menjadi dua komponen utama: pembentukan fitur meta dari prediksi model dasar, dan klasifikasi tingkat kedua oleh meta-learner.

3.3.5.1 Arsitektur Stacking

Arsitektur stacking yang digunakan dalam penelitian ini terdiri dari dua tingkat (*two-level architecture*) yang dirancang untuk menggabungkan kekuatan dari beberapa model dasar (*base learners*) secara adaptif. Pada tingkat pertama, tiga model CNN yaitu EfficientNetB0, MobileNetV2, dan MobileNetV3Small dilatih secara independen untuk melakukan klasifikasi citra X-ray ke dalam dua kelas, yaitu *Normal* dan *Pneumonia*. Ketiga model ini berfungsi sebagai *feature specialists* yang masing-masing mengekstraksi pola radiologis dari perspektif arsitektural yang berbeda. Setiap model menghasilkan keluaran berupa vektor probabilitas dari lapisan Softmax dengan format $P(\text{Normal}), P(\text{Pneumonia})$.

Keluaran probabilitas dari ketiga base learner kemudian dikombinasikan menjadi satu vektor baru berukuran enam elemen. Vektor ini disebut sebagai *meta-features*, karena merepresentasikan “opini probabilistik” kolektif dari seluruh model dasar. Alih-alih melakukan voting sederhana seperti pada teknik bagging, stacking menggunakan meta-features ini sebagai input untuk model tingkat kedua, yaitu *meta-learner* berupa Deep Neural Network (DNN). Meta-learner bertugas mempelajari pola interaksi antar prediksi model dasar, misalnya pada kondisi ketika satu model ragu-ragu sementara dua model sangat yakin, atau ketika semua model memberikan keyakinan rendah yang konsisten.



Gambar 3. 11 Arsitektur Stacking Ensemble

Dengan arsitektur dua tingkat ini, proses ensemble tidak hanya menggabungkan hasil prediksi secara statis, tetapi membangun mekanisme pengambilan keputusan yang lebih cerdas dan kontekstual. Stacking memungkinkan sistem untuk “*belajar dari hubungan antar model*”, bukan hanya memilih model yang paling dominan. Secara intuitif, pendekatan ini menyerupai proses diskusi antar tiga radiolog yang memberikan opini masing-masing, lalu seorang konsultan senior (*meta-learner*) mengevaluasi bobot keyakinan mereka sebelum menentukan diagnosis akhir.

3.3.5.2 Pembentukan *Meta-Features* dari Output Base Learners

Pada tahap ini, keluaran dari masing-masing *base learner* tidak langsung digunakan sebagai hasil klasifikasi akhir, melainkan dikonversi terlebih dahulu menjadi *meta-features* yang akan digunakan sebagai input bagi *meta-learner*. Setiap backbone CNN, yaitu EfficientNetB0, MobileNetV2, dan MobileNetV3Small, menghasilkan keluaran berupa vektor probabilitas Softmax terhadap dua kelas target [Normal, Pneumonia], Secara matematis, prediksi dari masing-masing model dasar dapat dituliskan sebagai:

$$f_1(x) = [p_1(0), p_1(1)], f_2(x) = [p_2(0), p_2(1)], f_3(x) = [p_3(0), p_3(1)] \quad (3.6)$$

dengan $p_i(0)$ menyatakan probabilitas kelas Normal dan $p_i(1)$ menyatakan probabilitas kelas Pneumonia dari model ke- i terhadap citra masukan x .

Seluruh probabilitas dari ketiga model kemudian digabungkan secara berurutan menjadi satu vektor fitur baru berukuran enam elemen:

$$z = [p_1(0), p_1(1), p_2(0), p_2(1), p_3(0), p_3(1)] \quad (3.7)$$

Vektor z ini disebut sebagai *meta-feature vector*, karena tidak lagi merepresentasikan fitur visual dari citra seperti pada lapisan konvolusi, melainkan *representasi keyakinan kolektif dari para base learner*. Dengan kata lain, setiap backbone dapat dipandang sebagai “*radiolog digital*” yang memberikan opini probabilistik, dan seluruh opini tersebut kemudian dikompilasi menjadi satu data baru yang siap dipelajari oleh *meta-learner*.

Pendekatan ini memberikan fleksibilitas lebih besar dibanding metode ensemble konvensional seperti *majority voting* atau *averaging*, karena meta-learner tidak hanya mempertimbangkan probabilitas tertinggi, tetapi juga *pola ketidaksepakatan antar model*. Misalnya, apabila EfficientNetB0 dan MobileNetV3Small menunjukkan probabilitas tinggi pada kelas Pneumonia, sementara MobileNetV2 memberikan prediksi ambigu, maka meta-learner dapat mempelajari bahwa pola ketidakseimbangan tersebut justru merupakan indikator kuat untuk menguatkan prediksi ke salah satu kelas.

Dengan demikian, tahap pembentukan meta-features merupakan fondasi utama yang memungkinkan stacking ensemble belajar secara adaptif dari kombinasi model, bukan sekadar melakukan agregasi statis. Tahap ini juga memastikan bahwa keputusan akhir tidak hanya ditentukan oleh satu model dominan, tetapi oleh *konsensus probabilistik* yang diolah kembali secara terstruktur oleh meta-learner pada tahap berikutnya.

3.3.5.3 Perancangan Meta-Learner

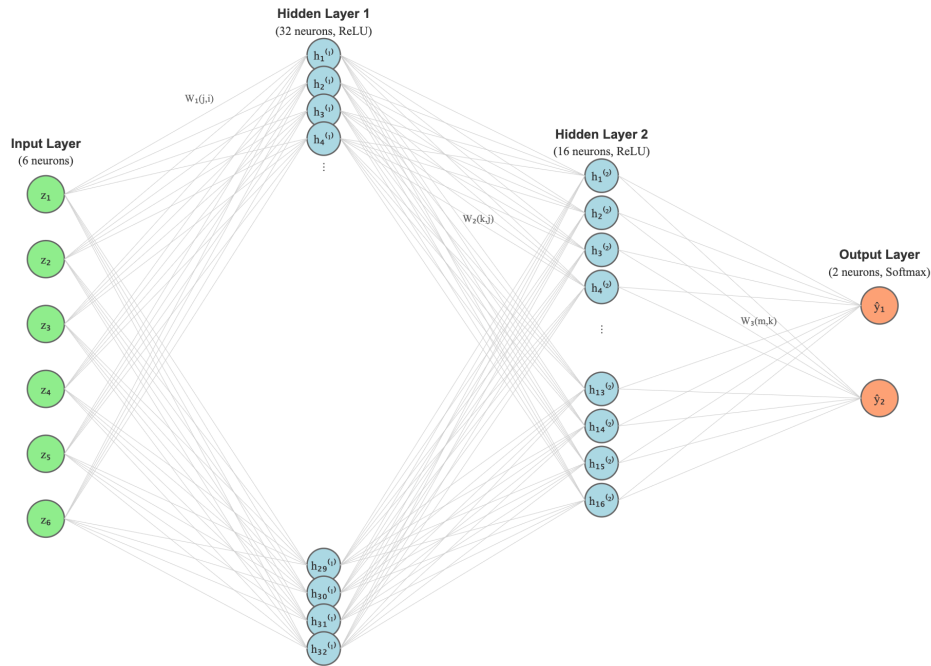
Meta-learner pada penelitian ini dirancang untuk menggabungkan *meta-features* berupa probabilitas keluaran dari tiga backbone CNN (EfficientNetB0, MobileNetV2, dan MobileNetV3Small). Setiap backbone menghasilkan dua probabilitas kelas [P(Normal),P(Pneumonia)] sehingga total input meta-learner adalah enam dimensi:

$$z = [p_1(0), p_1(1), p_2(0), p_2(1), p_3(0), p_3(1)] \in R^6 \quad (3.7)$$

Yang mana z adalah vektor fitur hasil penggabungan (concatenation) dari seluruh probabilitas prediksi yang dikeluarkan oleh tiga model dasar.

- $p_1(0)$: probabilitas bahwa citra termasuk kelas **Normal** berdasarkan prediksi EfficientNetB0.
- $p_1(1)$: probabilitas bahwa citra termasuk kelas **Pneumonia** berdasarkan prediksi EfficientNetB0.
- $p_2(0)$: probabilitas bahwa citra termasuk kelas **Normal** berdasarkan prediksi MobileNetV2.
- $p_2(1)$: probabilitas bahwa citra termasuk kelas **Pneumonia** berdasarkan prediksi MobileNetV2.
- $p_3(0)$: probabilitas bahwa citra termasuk kelas **Normal** berdasarkan prediksi MobileNetV3Small.
- $p_3(1)$: probabilitas bahwa citra termasuk kelas **Pneumonia** berdasarkan prediksi MobileNetV3Small.

Arsitektur *meta-learner* menggunakan *Dense Neural Network (DNN)* dengan dua lapisan tersembunyi (*hidden layers*). Pemilihan arsitektur ini didasarkan pada pertimbangan bahwa dimensi input relatif kecil, sehingga cukup dengan dua *hidden layer* untuk memungkinkan model belajar transformasi non-linear tanpa meningkatkan risiko *overfitting* secara berlebihan.



Gambar 3. 12 Arsitektur *Meta Learner (Dense Neural Network)*

Struktur meta-learner adalah sebagai berikut:

1. Lapisan Input: menerima vektor meta-feature z .
2. Hidden Layer 1: 32 unit neuron dengan fungsi aktivasi ReLU,

$$h^{(1)} = \text{ReLU}(W_1 z + b_1) \quad (3.8)$$

3. Hidden Layer 2: 16 unit neuron dengan fungsi aktivasi ReLU,

$$h^{(2)} = \text{ReLU}(W_2 h^{(1)} + b_2) \quad (3.9)$$

4. Lapisan Output: 2 unit neuron dengan fungsi aktivasi Softmax untuk menghasilkan probabilitas akhir dua kelas:

$$\hat{y} = \text{Softmax}(W_3 h^{(2)} + b_3) \quad (3.10)$$

Dengan

$$\text{Softmax}(z)_j = \frac{e^{z_j}}{\sum_{k=1}^2 e^{z_k}} \quad (3.11)$$

Secara umum, formulasi feedforward pada DNN ini dapat ditulis:

$$a^{(l)} = f(W^{(l)} a^{(l-1)} + b^{(l)}), \quad l = 1, 2 \quad (3.12)$$

$$\hat{y} = \text{Softmax}(W^{(3)} a^{(2)} + b^{(3)}) \quad (3.13)$$

dengan $a(0) = z$ sebagai input meta-feature, dan fungsi aktivasi hidden layer menggunakan ReLU.

Penggunaan probabilitas Softmax sebagai input dipilih karena lebih interpretatif dibandingkan logit mentah. Dengan begitu, meta-learner belajar langsung dari tingkat keyakinan setiap backbone, bukan hanya nilai skor mentah. Dengan desain dua hidden layer ini, meta-learner berperan sebagai pengambil keputusan kolektif yang mempelajari kombinasi optimal dari probabilitas base learners. Struktur ini memungkinkan sistem mengenali pola kompleks, seperti dominasi satu backbone dengan keyakinan tinggi atau kombinasi moderat dari beberapa backbone, sehingga menghasilkan prediksi akhir yang lebih stabil dan akurat.

3.4 Implementasi Sistem

Implementasi sistem stacking ensemble pada penelitian ini dilakukan secara bertahap dan terstruktur. Proses dimulai dari persiapan data, yaitu pemanggilan dataset, penyeragaman ukuran citra, normalisasi nilai piksel, dan pembagian data ke dalam tiga subset utama: training set, validation set, dan testing set. Setelah data siap, tiga model dasar (EfficientNetB0, MobileNetV2, dan MobileNetV3Small) dilatih secara independen untuk menghasilkan probabilitas prediksi dua kelas, yaitu Normal dan Pneumonia. Probabilitas keluaran dari seluruh base learners kemudian digabungkan secara horizontal sehingga membentuk meta-features berdimensi enam pada setiap sampel. Vektor meta-features ini selanjutnya digunakan sebagai masukan bagi meta-learner, yaitu jaringan saraf tiruan sederhana dengan dua lapisan tersembunyi. Meta-learner dilatih untuk mempelajari pola hubungan antar keluaran model dasar agar mampu menghasilkan keputusan klasifikasi akhir secara lebih stabil dan akurat.

Pada tahap pengujian, citra baru diproses melalui alur yang sama: base learners memberikan probabilitas awal, probabilitas tersebut digabungkan menjadi meta-features, dan meta-learner menentukan hasil klasifikasi akhir. Dengan pendekatan ini, keputusan tidak bergantung pada satu model tunggal, melainkan merupakan hasil pertimbangan kolektif dari seluruh base learners melalui proses pembelajaran tingkat kedua.

BAB IV

UJI COBA DAN PEMBAHASAN

4.1 Skenario Pengujian

Langkah pertama sebelum melakukan pengujian adalah mempersiapkan data yang akan digunakan dalam penelitian. Seperti yang telah dijelaskan pada Sub-bab 3.2, dataset yang digunakan bersumber dari Chest X-Ray Images (Pneumonia) yang tersedia pada platform Kaggle. Dataset ini terdiri dari dua kelas, yaitu Normal dan Pneumonia, dengan total 5.863 citra.

Untuk tahap pengujian, penelitian ini menggunakan subset data uji (test set) yang berjumlah 624 citra, dengan rincian 234 citra kelas Normal dan 390 citra kelas Pneumonia. Subset ini dipilih karena tidak digunakan dalam proses pelatihan maupun validasi, sehingga mampu merepresentasikan kemampuan generalisasi model secara objektif.

Sebelum masuk ke tahap pengujian, seluruh citra dalam subset uji telah melalui proses preprocessing sesuai yang dijelaskan pada Bab III, meliputi resizing citra menjadi 224×224 piksel, serta normalisasi nilai piksel. Hasil preprocessing ini kemudian siap untuk digunakan dalam proses ekstraksi fitur dan klasifikasi pada model dasar maupun stacking ensemble.

Pada tahap stacking, keluaran tiap model dasar berupa probabilitas dua kelas (NORMAL, PNEUMONIA). Untuk tiga backbone (EfficientNetB0, MobileNetV2, MobileNetV3Small), vektor fitur yang masuk ke meta-learner adalah hasil concatenate enam probabilitas:

$$\mathbf{z} = \left[p_{\text{Eff}}^{(0)}, p_{\text{Eff}}^{(1)}, p_{V2}^{(0)}, p_{V2}^{(1)}, p_{V3}^{(0)}, p_{V3}^{(1)} \right] \in R^6 \quad (4.1)$$

dengan indeks (0) untuk NORMAL dan (1) untuk PNEUMONIA.

Meta-learner diimplementasikan sebagai MLP dengan arsitektur tetap:

$$\text{Input}(6) \rightarrow \text{Dense}(64) \xrightarrow{\text{BN}} \text{ReLU} \xrightarrow{\text{Dropout } 0.3} \text{Dense}(32) \xrightarrow{\text{BN}} \text{ReLU} \xrightarrow{\text{Dropout } 0.3} \text{Dense}(2) \xrightarrow{\text{Softmax}} .$$

Bobot diinisialisasi dengan He initialization. Terdapat Batch Normalization (BN) pada dua hidden layer aktivasi ReLU, serta Dropout 0,3.

Forward pass (training)

Untuk setiap lapisan $\mathbf{l}$ yang bertipe Dense (affine):

$$\mathbf{a}^{(1)} = \mathbf{z} W_1 + \mathbf{b}_1, \quad \tilde{\mathbf{a}}^{(1)} = \text{BN}(\mathbf{a}^{(1)}) = \gamma_1 \odot \frac{\mathbf{a}^{(1)} - \mu_1}{\sqrt{\sigma_1^2 + \varepsilon}} + \beta_1, \quad (4.2)$$

$$\mathbf{h}^{(1)} = \text{ReLU}(\tilde{\mathbf{a}}^{(1)}), \quad \hat{\mathbf{h}}^{(1)} = \text{Dropout}(\mathbf{h}^{(1)}), \quad (4.3)$$

$$\mathbf{a}^{(2)} = \hat{\mathbf{h}}^{(1)} W_2 + \mathbf{b}_2, \quad \tilde{\mathbf{a}}^{(2)} = \text{BN}(\mathbf{a}^{(2)}) = \gamma_2 \odot \frac{\mathbf{a}^{(2)} - \mu_2}{\sqrt{\sigma_2^2 + \varepsilon}} + \beta_2, \quad (4.4)$$

$$\mathbf{h}^{(2)} = \text{ReLU}(\tilde{\mathbf{a}}^{(2)}), \quad \hat{\mathbf{h}}^{(2)} = \text{Dropout}(\mathbf{h}^{(2)}), \quad (4.5)$$

$$\mathbf{o} = \hat{\mathbf{h}}^{(2)} W_3 + \mathbf{b}_3, \quad \hat{\mathbf{y}} = \text{softmax}(\mathbf{o}), \quad \hat{y}_k = \frac{e^{o_k}}{\sum_{j=1}^2 e^{o_j}}. \quad (4.6)$$

Fungsi Cost dan regularisasi

Meta-learner mengoptimasi **cross-entropy** dua kelas dengan **regularisasi L2** pada bobot Dense:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^N \sum_{k=1}^2 y_{ik} \log \hat{y}_{ik} + \frac{\lambda}{2} (\|W_1\|_2^2 + \|W_2\|_2^2 + \|W_3\|_2^2), \quad (4.7)$$

dengan $\mathbf{y}_{ik}$ adalah label *one-hot*, $\lambda = 10^{-4}$ (*weight decay*).

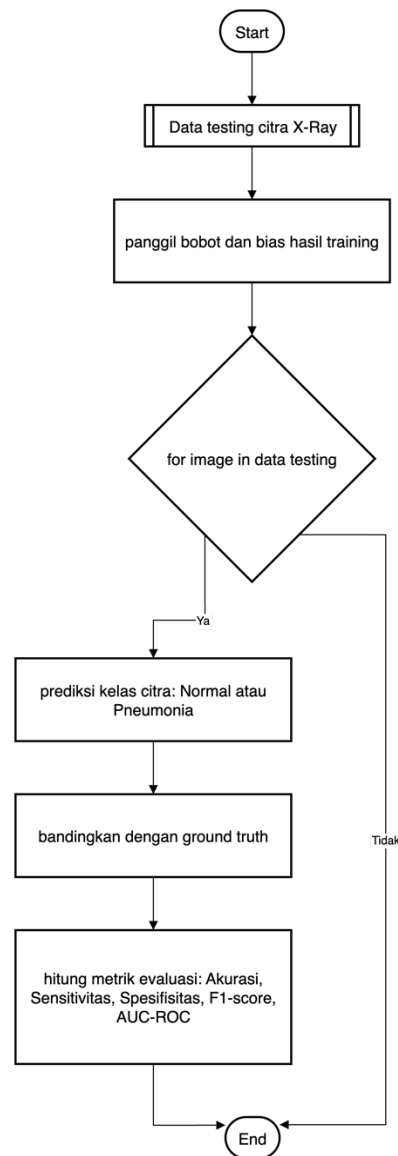
Backpropagation

Gradien pada lapisan keluaran *softmax-cross-entropy*:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{o}} = \hat{\mathbf{y}} - \mathbf{y}, \quad \frac{\partial \mathcal{L}}{\partial \mathbf{W}_3} = \hat{\mathbf{h}}^{(2)\top} \frac{\partial \mathcal{L}}{\partial \mathbf{o}} + \lambda \mathbf{W}_3, \quad \frac{\partial \mathcal{L}}{\partial \mathbf{b}_3} = \text{sum}_{\text{batch}} \left(\frac{\partial \mathcal{L}}{\partial \mathbf{o}} \right). \quad (4.8)$$

Gradien kemudian dirambatkan ke belakang melewati **Dropout** (identitas pada mask), **ReLU** ($1(\cdot > 0)$), dan **BatchNorm** (turunan terhadap γ, β , mean, dan var batch) untuk mendapatkan turunan $\mathbf{W}_2, \mathbf{b}_2, \gamma_2, \beta_2$ dan $\mathbf{W}_1, \mathbf{b}_1, \gamma_1, \beta_1$. Komponen L2 menambah $\lambda \mathbf{W}_1$ ada tiap $\partial \mathcal{L} / \partial \mathbf{W}_1$.

Setelah model melalui tahap pelatihan (training) untuk mempelajari pola dari data latih yang diberikan, langkah selanjutnya adalah tahap pengujian (testing). Tujuan utama dari tahap ini adalah untuk mengevaluasi kinerja model dalam melakukan klasifikasi citra X-Ray paru-paru yang belum pernah dilihat sebelumnya. Dengan demikian, performa model dapat dinilai secara objektif dan menunjukkan kemampuan generalisasi terhadap data baru.



Gambar 4. 1 Flowchart Pengujian

Pada penelitian ini, pengujian dilakukan terhadap tiga model tunggal, yaitu EfficientNetB0, MobileNetV2, dan MobileNetV3Small, serta satu model stacking ensemble. Setiap model dievaluasi menggunakan dua skema pembagian data, yaitu 80% data latih – 20% data uji dan 70% data latih – 30% data uji. Untuk memperoleh hasil evaluasi yang lebih stabil dan tidak bergantung pada satu pembagian data tertentu, seluruh skenario pengujian menerapkan metode validasi silang K-Fold

dengan nilai $K = 5$.

Pada skema pembagian data 70% data latih dan 30% data uji, dataset terlebih dahulu dibagi secara tetap menjadi data latih dan data uji. Selanjutnya, metode K-Fold Cross Validation diterapkan hanya pada data latih, di mana data latih tersebut dibagi menjadi lima fold yang saling tidak tumpang tindih. Pada setiap iterasi, empat fold digunakan untuk proses pelatihan model dan satu fold digunakan sebagai data validasi internal. Proses ini diulang hingga setiap fold pernah digunakan satu kali sebagai data validasi. Sementara itu, data uji sebesar 30% dipisahkan sepenuhnya dan tidak dilibatkan dalam proses pelatihan maupun validasi K-Fold, melainkan hanya digunakan untuk evaluasi akhir model. Dengan pendekatan ini, tidak terjadi tumpang tindih data antara data latih dan data uji, sehingga risiko kebocoran data dapat dihindari serta hasil evaluasi model menjadi lebih reliabel dan objektif.

Karena dataset penelitian ini memiliki ketidakseimbangan kelas antara Normal dan Pneumonia, proses pelatihan juga melibatkan teknik oversampling berbasis SMOTE untuk menyeimbangkan distribusi data pada kelas minoritas. Dengan pendekatan ini, model CNN dapat belajar pola secara lebih stabil tanpa bias berlebihan terhadap kelas mayoritas. Seluruh proses SMOTE dilakukan hanya pada data latih untuk menghindari kebocoran data.

Dalam penelitian klasifikasi medis, distribusi kelas yang tidak seimbang (imbalanced data) merupakan masalah umum, terutama ketika jumlah citra Pneumonia lebih banyak dibandingkan citra Normal. Ketidakseimbangan ini dapat menyebabkan model condong mempelajari pola mayoritas (bias ke kelas

Pneumonia) sehingga menurunkan sensitivitas maupun spesifisitas. Untuk mengatasi masalah ini, penelitian ini menggunakan pendekatan Synthetic Minority Over-sampling Technique (SMOTE), yaitu metode oversampling berbasis sintesis yang bekerja dengan membangkitkan sampel baru pada kelas minoritas melalui interpolasi linier antar tetangga terdekat (k-nearest neighbors).

Berbeda dengan random oversampling yang hanya menggandakan citra secara duplikasi sehingga berpotensi menyebabkan overfitting, SMOTE membentuk titik fitur baru di ruang feature-space, sehingga variasi data meningkat dan model memperoleh representasi kelas minoritas yang lebih kaya. Secara formal, jika x_i adalah sampel kelas minoritas dan x_{NN} adalah tetangga terdekatnya, maka SMOTE menghasilkan data sintesis:

$$\tilde{x} = x_i + \alpha(x_{NN} - x_i), \quad \alpha \sim U(0, 1) \quad (4.9)$$

Dengan cara ini, distribusi data pada kelas minoritas menjadi lebih seimbang tanpa mengubah karakteristik asli data secara ekstrem. Peningkatan variasi feature-space ini terbukti membantu model CNN belajar pola representatif secara lebih stabil.

Namun, SMOTE juga memiliki batasan. Karena sintesis dilakukan di ruang fitur dan bukan pada citra asli, SMOTE dapat menghasilkan representasi yang tidak sepenuhnya mencerminkan struktur kompleks pada citra X-Ray. Selain itu, penggunaan SMOTE harus dipastikan hanya diterapkan pada data latih dan **tidak pernah** pada data validasi maupun test set, untuk menghindari kebocoran data (data leakage). Oleh karena itu, dalam penelitian ini SMOTE diterapkan secara ketat pada

bagian data latih setelah proses splitting 80:20 maupun 70:30, dan tidak pernah diterapkan pada subset uji 624 citra.

Langkah awal dalam pengujian adalah menyiapkan data uji yang dipisahkan dari data latih. Selanjutnya, model memanggil bobot hasil pelatihan dan digunakan untuk menghasilkan prediksi kelas citra. Prediksi yang dihasilkan kemudian dibandingkan dengan label sebenarnya (ground truth). Dari hasil perbandingan ini, dihitung nilai metrik evaluasi berupa akurasi, sensitivitas, spesifisitas, F1-score, dan AUC-ROC, yang diperoleh melalui confusion matrix.

Untuk memberikan gambaran lebih sistematis, kombinasi skenario pengujian yang digunakan dalam penelitian ini ditunjukkan pada Tabel 4.1 berikut.

Tabel 4. 1 Kombinasi Skenario Pengujian

No.	Kode Skenario	Model	Dataset	Pembagian Data	Metode Evaluasi
1	1A-K	EfficientNetB0	Subset ($\pm 1k/class$)	80:20	K-Fold (K=5)
2	1B-K	EfficientNetB0	Subset ($\pm 1k/class$)	70:30	K-Fold (K=5)
3	2A-K	MobileNetV2	Subset ($\pm 1k/class$)	80:20	K-Fold (K=5)
4	2B-K	MobileNetV2	Subset ($\pm 1k/class$)	70:30	K-Fold (K=5)
5	3A-K	MobileNetV3Small	Subset ($\pm 1k/class$)	80:20	K-Fold (K=5)
6	3B-K	MobileNetV3Small	Subset ($\pm 1k/class$)	70:30	K-Fold (K=5)
No.	Kode Skenario	Model	Dataset	Pembagian Data	Metode Evaluasi

7	4A-K	Stacking Ensemble	Subset ($\pm 1k/class$)	80:20	K-Fold (K=5)
8	4B-K	Stacking Ensemble	Subset ($\pm 1k/class$)	70:30	K-Fold (K=5)
9	1A-F	EfficientNetB0	Full	80:20	K-Fold (K=5)
10	1B-F	EfficientNetB0	Full	70:30	K-Fold (K=5)
11	2A-F	MobileNetV2	Full	80:20	K-Fold (K=5)
12	2B-F	MobileNetV2	Full	70:30	K-Fold (K=5)
13	3A-F	MobileNetV3Small	Full	80:20	K-Fold (K=5)
14	3B-F	MobileNetV3Small	Full	70:30	K-Fold (K=5)
15	4A-F	Stacking Ensemble	Full	80:20	K-Fold (K=5)
16	4B-F	Stacking Ensemble	Full	70:30	K-Fold (K=5)
17	SMOTE_1A-F	EfficientNetB0	Full + Smote	80:20	K-Fold(5)
18	SMOTE_4A-F	Stacking Ensemble	Full + Smote	80:20	K-Fold(5)
19	SMOTE_1B-F	EfficientNetB0	Full + Smote	70:30	K-Fold(5)
20	SMOTE_4B-F	Stacking Ensemble	Full + Smote	70:30	K-Fold(5)

Kode eksperimen dalam penelitian ini digunakan untuk merepresentasikan jenis model dan skema pembagian data yang diterapkan. Angka 1 hingga 3 menunjukkan penggunaan model tunggal, di mana angka 1 merepresentasikan EfficientNetB0, angka 2 merepresentasikan MobileNetV2, dan angka 3 merepresentasikan MobileNetV3Small. Angka 4 menunjukkan penggunaan model *stacking ensemble* yang merupakan gabungan dari ketiga model dasar tersebut dengan bantuan *meta-learner*. Huruf A pada kode menunjukkan pembagian data dengan rasio 80:20 antara data latih dan data uji, sedangkan huruf B menunjukkan

pembagian data dengan rasio 70:30. Huruf K menandakan bahwa proses evaluasi dilakukan menggunakan metode validasi silang K-Fold dengan jumlah lipatan sebanyak lima. Huruf F menunjukkan bahwa seluruh dataset digunakan dalam proses pelatihan dan pengujian (*full dataset*), sementara kode tanpa huruf F menunjukkan bahwa eksperimen dilakukan menggunakan subset data dengan jumlah sekitar 1000 citra per kelas.

Dengan skenario di atas, penelitian ini menghasilkan delapan kombinasi pengujian yang mencakup baik model tunggal maupun stacking ensemble, sehingga dapat dilakukan perbandingan performa secara menyeluruh.

Sebelum masuk kedalam pembahasan setiap skenario, nilai dari parameter-parameter harus ditentukan terlebih dahulu. Setiap skenario akan memiliki nilai parameter yang sama, lebih detail mengenai nilai setiap parameter dijelaskan di dalam Tabel 4.2

Tabel 4. 2 Nilai masing-masing parameter

Parameter	Nilai	Keterangan
Random Seed	42	Digunakan untuk reproduibilitas
Jumlah Fold (K)	5	K-Fold Cross Validation
Epoch Base Model	30	Jumlah epoch untuk pelatihan model dasar
Epoch Meta Model	30	Jumlah epoch untuk pelatihan meta-learner
Batch Size Base	64	Ukuran batch saat pelatihan model dasar
Batch Size Meta	64	Ukuran batch saat pelatihan meta-learner

Evaluasi performa model dilakukan secara kuantitatif berdasarkan metrik yang dihitung dari *confusion matrix* serta kurva evaluasi tambahan. Lima indikator utama digunakan dalam penelitian ini, yaitu akurasi, sensitivitas, spesifisitas, F1-Score, dan AUC-ROC.

Dalam penelitian ini, beberapa metrik evaluasi digunakan untuk menilai

performa model deteksi pneumonia. Akurasi digunakan untuk mengukur proporsi prediksi yang benar dari seluruh data uji, sehingga memberikan gambaran umum mengenai kinerja model. Namun, dalam konteks medis, akurasi saja tidak cukup. Oleh karena itu, sensitivitas (True Positive Rate) menjadi metrik yang diprioritaskan karena menunjukkan kemampuan model dalam mendeteksi kasus pneumonia secara benar. Metrik ini sangat penting mengingat kesalahan dalam mengidentifikasi pasien yang sebenarnya sakit (false negative) dapat menimbulkan risiko serius bagi keselamatan pasien.

Di sisi lain, spesifisitas (True Negative Rate) digunakan untuk menilai kemampuan model dalam mengenali pasien Normal secara tepat, sehingga mengurangi kemungkinan kesalahan diagnosis pada individu sehat (false positive). Selain itu, F1-Score, sebagai rata-rata harmonik antara presisi dan sensitivitas, dipilih untuk memberikan penilaian yang lebih seimbang pada dataset yang berpotensi tidak seimbang. Metrik ini memastikan bahwa model tidak hanya mampu mendeteksi kasus positif, tetapi juga menjaga konsistensi performa antara precision dan recall.

Selanjutnya, AUC-ROC (Area Under the Curve) digunakan untuk mengevaluasi kemampuan model dalam membedakan kelas positif dan negatif pada berbagai nilai ambang keputusan (threshold). Nilai AUC yang tinggi menunjukkan bahwa model mampu memisahkan pasien pneumonia dan pasien Normal secara konsisten di berbagai kondisi threshold. Dengan kombinasi metrik-metrik tersebut, evaluasi performa model menjadi lebih komprehensif dan sesuai dengan kebutuhan

analisis dalam bidang kesehatan. Secara matematis, metrik evaluasi tersebut dirumuskan sebagai berikut:

$$\text{Akurasi} = \frac{TP + TN}{TP + TN + FP + FN} \quad (4.9)$$

$$\text{Sensitivitas} = \frac{TP}{TP + FN} \quad (4.10)$$

$$\text{Spesifisitas} = \frac{TN}{TN + FP} \quad (4.11)$$

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4.12)$$

Dengan definisi Precision dan Recall:

$$\text{Precision} = \frac{TP}{TP + FP} \quad \text{dan} \quad \text{Recall} = \frac{TP}{TP + FN} \quad (4.13)$$

Versi langsung dari confusion matrix:

$$F1 = 2 \times \frac{TP}{2TP + FP + FN} \quad (4.14)$$

AUC ROC:

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR}) \, d(\text{FPR}) \quad (4.15)$$

Dengan:

$$\text{TPR} = \frac{TP}{TP + FN} \quad \text{dan} \quad \text{FPR} = \frac{FP}{FP + TN} \quad (4.16)$$

dengan:

- a. *True Positive (TP)*: jumlah citra pneumonia yang berhasil diklasifikasikan dengan benar,
- b. *False Negative (FN)*: jumlah citra pneumonia yang salah diklasifikasikan sebagai Normal,
- c. *True Negative (TN)*: jumlah citra Normal yang diklasifikasikan dengan benar,
- d. *False Positive (FP)*: jumlah citra Normal yang salah diklasifikasikan sebagai pneumonia.

Untuk memastikan hasil evaluasi tidak bergantung pada satu skema pembagian data saja, dua pendekatan validasi diterapkan dalam eksperimen ini, yaitu:

- Train-Test Split dengan beberapa skenario (misalnya 80:20 dan 70:30) untuk melihat konsistensi performa model pada proporsi data yang berbeda.
- K-Fold Cross Validation ($K = 5$) untuk memperoleh rata-rata performa yang lebih robust dan mengurangi bias akibat pemilihan subset data tertentu.

Dengan skema evaluasi ini, performa model stacking ensemble dapat dibandingkan secara objektif dengan model tunggal baik dari sisi akurasi maupun kestabilan hasil prediksi.

4.2 Hasil Pengujian

4.2.1 Hasil Pengujian Model Tunggal: EfficientNetB0

EfficientNetB0 dipilih sebagai salah satu model tunggal dalam penelitian ini karena arsitektur compound scaling yang mampu menyeimbangkan kedalaman, lebar, dan resolusi jaringan secara efisien. Model ini diharapkan dapat memberikan performa tinggi dengan kompleksitas komputasi yang relatif lebih rendah

dibandingkan arsitektur yang lebih besar.

Pengujian dilakukan pada empat scenario , yaitu 80:20 (1A-K), 70:30 (1B-K), 80:20 (1A-F), dan 70:30 (1B-F) menggunakan validasi silang K-Fold dengan K = 5. Hasil pengujian ditunjukkan pada Tabel 4.3.

Tabel 4. 3 Hasil Evaluasi Model EfficientNetB0

Kode	Split	Akurasi	Presisi	Sensitivitas	Spesifisitas	F1-Score	AUC-ROC
1A-K	80:20	97,75%	97,28%	98,62%	96,70%	97,95%	99,76%
1B-K	70:30	97,50%	97,47%	97,78%	97,18%	97,63%	99,73%
1A-F	80:20	96,93%	100,0%	95,97%	100,0%	97,95%	96,93%
1B-F	70:30	97,57%	97,47%	97,78%	97,18%	97,63%	99,73%

Interpretasi Skenario 1A-K (Split 80:20)

Dari log hasil training skenario **1A-K** (EfficientNetB0, 80:20, K=5), tabel hasil tiap fold bisa dilihat pada table 4.4 berikut:

Tabel 4. 4 Log Training Skenario 1A-K Fold = 5

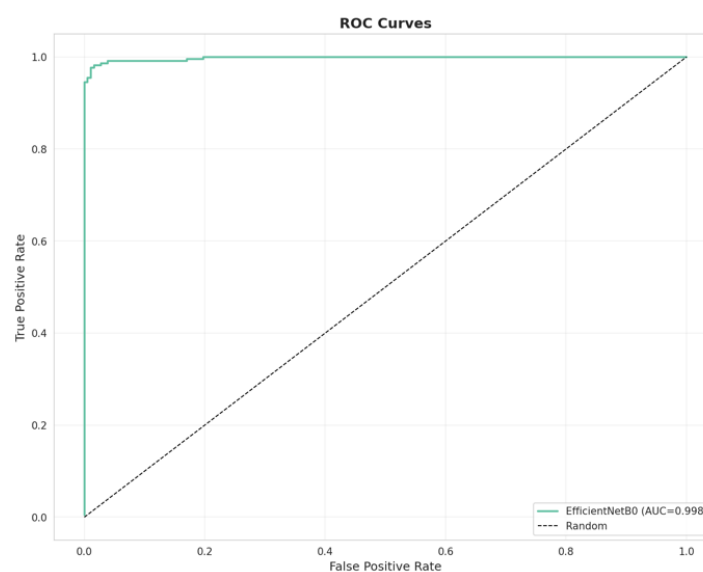
Fold	Val Accuracy
1	94,36%
2	91,85%
3	94,67%
4	96,87%
5	95,06%
Rata-rata (OOF)	95,44%

Berdasarkan hasil evaluasi, model menunjukkan performa yang sangat baik dalam mendeteksi pneumonia. Hal ini terlihat dari nilai **akurasi sebesar 97,75%** dengan **sensitivitas 98,62%**, yang menandakan model mampu mengidentifikasi kasus pneumonia secara tepat dengan sangat sedikit kesalahan. Nilai **spesifisitas 96,70%** juga tergolong tinggi, sehingga risiko salah mendiagnosis pasien sehat sebagai sakit (false positive) relatif kecil.

Analisis lebih lanjut melalui **confusion matrix** memperlihatkan bahwa dari total 400 sampel, model mampu menghasilkan **215 True Positive** dan **176 True Negative**, dengan hanya **6 False Positive** dan 3 False Negative. Proporsi kesalahan yang sangat kecil ini menunjukkan konsistensi model dalam membedakan citra pneumonia dan normal.

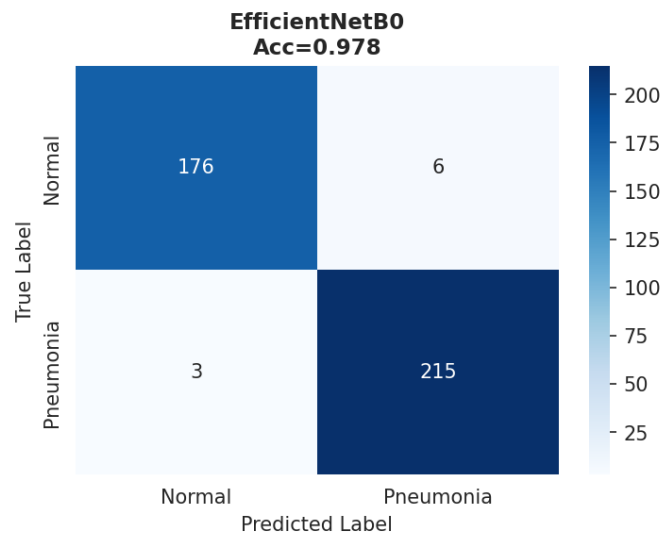
Selain itu, ROC Curve (Gambar 4.2) menghasilkan nilai AUC sebesar 0,9976, di mana kurva hampir menempel pada sudut kiri atas. Hal ini mengindikasikan kemampuan pemisahan kelas yang mendekati sempurna pada berbagai nilai threshold. Sementara itu, Precision–Recall Curve (Gambar 4.4) memberikan nilai PR-AUC sebesar 0,9929, menunjukkan bahwa performa model tetap stabil meskipun terjadi variasi distribusi antara kelas positif dan negatif.

Secara keseluruhan, hasil ini mengonfirmasi bahwa model memiliki kinerja yang sangat kuat, baik dalam mengenali kasus pneumonia maupun mempertahankan akurasi diagnosis pada pasien normal.



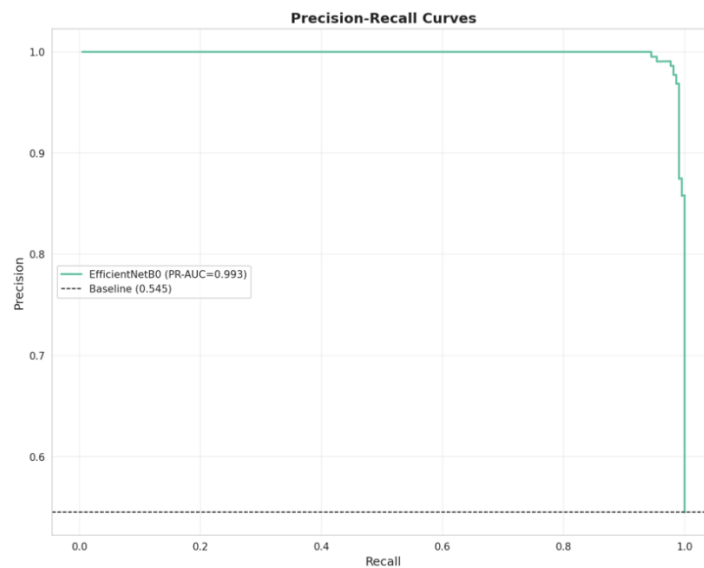
Gambar 4. 2 ROC Curve EfficientNetB0 (Skenario 1A-K)

Gambar 4.2 menampilkan kurva ROC pada skenario 1A-K (80:20) dengan AUC 0,9976. Kurva mendekati sudut kiri atas, menandakan performa klasifikasi yang hampir sempurna.



Gambar 4. 3 Confusion Matrix EfficientNetB0 (Skenario 1A-K)

Gambar 4.3 menunjukkan confusion matrix hasil prediksi pada skenario 1A-K, di mana hanya terdapat sedikit kesalahan klasifikasi (6 FP dan 3 FN dari total 400 sampel).



Gambar 4. 4 Precision-Recall Curve EfficientNetB0 (Skenario 1A-K)

Gambar 4.4 memperlihatkan kurva Precision-Recall pada skenario 1A-K, dengan area di bawah kurva (PR-AUC) sebesar 0,9929.

Interpretasi Skenario 1B-K (Split 70:30)

Dari log hasil training skenario **1B-K** (EfficientNetB0, 70:30, K=5), tabel hasil tiap fold bisa dilihat pada table 4.5 berikut:

Tabel 4. 5 Log Training Skenario 1B-K Fold = 5

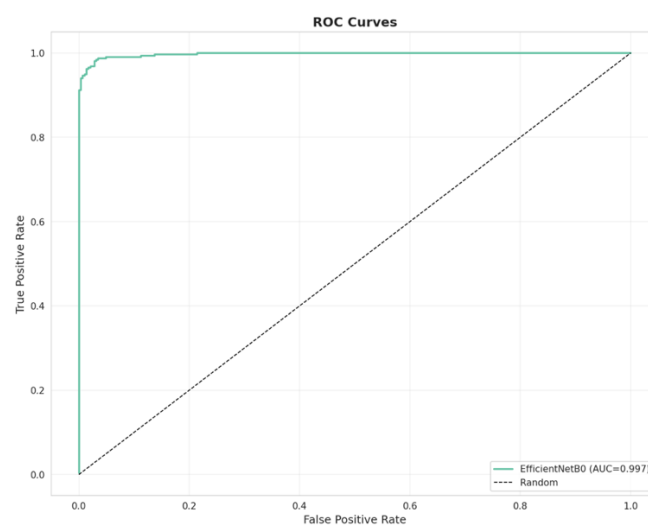
Fold	Val Accuracy
1	94,62%
2	94,98%
3	94,27%
4	96,42%
5	94,01%
Rata-rata (OOF)	95,36%

Pada skenario evaluasi ini, model menunjukkan performa yang tetap tinggi meskipun proporsi data uji diperbesar. Akurasi mencapai 97,50% dengan sensitivitas 97,78%, yang menandakan kemampuan model dalam mendeteksi kasus pneumonia tetap konsisten. Nilai spesifisitas sebesar 97,18% turut menunjukkan

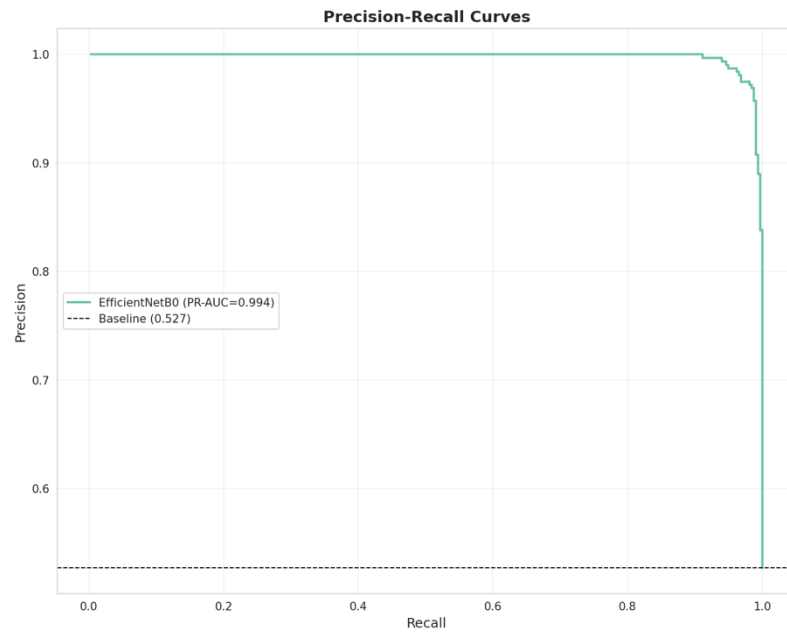
bahwa model mampu menekan jumlah false positive, meskipun jumlah sampel pengujian meningkat.

Hasil confusion matrix memperlihatkan bahwa dari total 600 sampel, model menghasilkan 309 true positive dan 276 true negative, dengan hanya 8 false positive dan 7 false negative. Distribusi ini mengindikasikan kestabilan performa model dalam membedakan antara citra pneumonia dan citra normal pada skenario dengan data uji yang lebih besar.

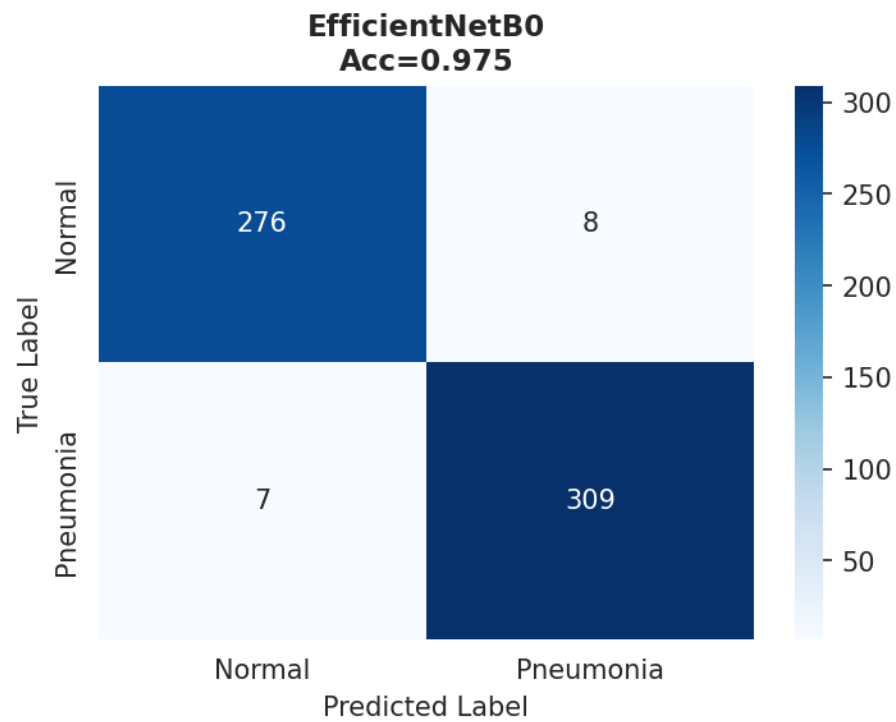
Selanjutnya, ROC Curve pada Gambar 4.5 menunjukkan nilai AUC sebesar 0,9973, yang hampir identik dengan skenario 1A-K, sehingga menguatkan bahwa model memiliki kemampuan pemisahan kelas yang sangat baik. Precision-Recall Curve pada Gambar 4.6 mencatat nilai PR-AUC sebesar 0,9941, yang bahkan sedikit lebih baik dibandingkan skenario sebelumnya. Secara keseluruhan, hasil-hasil tersebut menunjukkan bahwa model dapat mempertahankan performa yang sangat kuat meskipun jumlah data uji bertambah.



Gambar 4. 5 ROC Curve EfficientNetB0 (Skenario 1B-K)



Gambar 4. 6 Precision-Recall Curve EfficientNetB0 (Skenario 1B-K)



Gambar 4. 7 Confusion Matrix EfficientNetB0 (Skenario 1B-K)

Interpretasi Skenario 1A-F (Split 80:20, Full Dataset)

Dari log hasil training skenario 1A-F (EfficientNetB0, 80:20, K=5), tabel hasil tiap fold dapat dilihat pada Tabel 4.6 berikut:

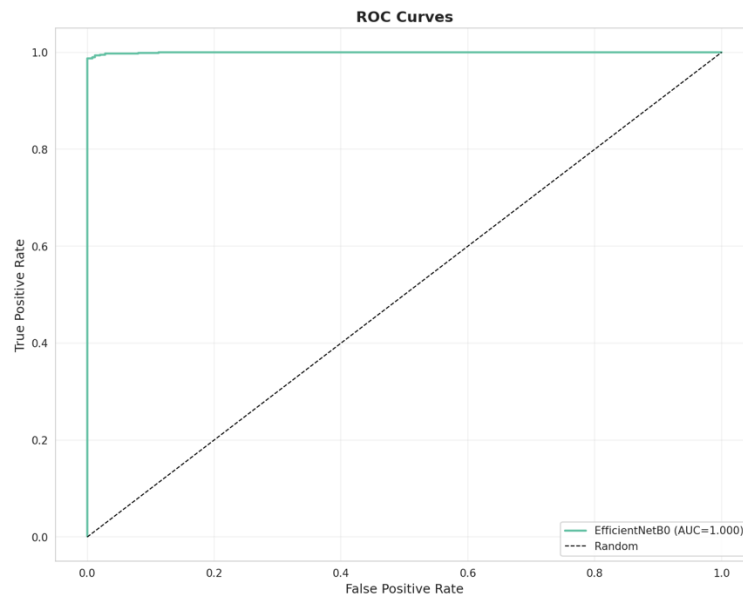
Tabel 4. 6 Log Training Skenario 1A-F Fold = 5

Fold	Val Accuracy
1	97,84%
2	97,96%
3	97,72%
4	97,36%
5	98,44%
Rata-rata (OOF)	97,44%

Hasil pengujian pada test set menunjukkan bahwa model memiliki performa yang sangat baik dalam mendeteksi pneumonia maupun mempertahankan ketepatan diagnosis pada pasien normal. Akurasi mencapai 96,93% dengan presisi 100,00%, sehingga seluruh prediksi positif yang dihasilkan model benar-benar merupakan kasus pneumonia. Sensitivitas sebesar 95,97% mengindikasikan bahwa sebagian besar pasien pneumonia berhasil dideteksi, meskipun masih terdapat sejumlah kasus yang terlewat. Spesifisitas berada pada nilai 100,00%, yang berarti model mampu mengidentifikasi seluruh pasien sehat tanpa menghasilkan satu pun false positive. Nilai F1-score sebesar 97,95% menunjukkan keseimbangan yang kuat antara presisi dan recall, sementara ROC-AUC sebesar 99,96% serta PR-AUC mencapai 99,85% menegaskan kemampuan model dalam memisahkan kelas positif dan negatif secara konsisten pada berbagai threshold.

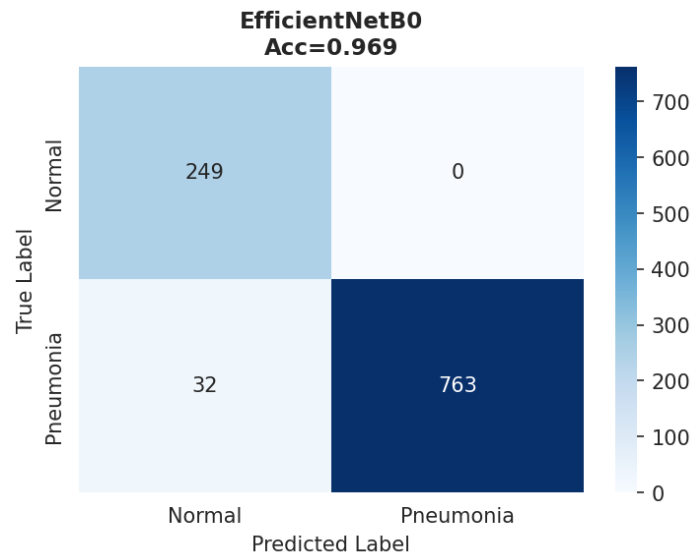
Confusion matrix memperlihatkan bahwa model menghasilkan 249 true negative dan tidak terdapat false positive, sehingga setiap pasien normal dapat diklasifikasikan dengan benar. Di sisi lain, terdapat 763 true positive dan 32 false negative, yang menunjukkan bahwa meskipun model sangat andal, masih terdapat sejumlah kasus pneumonia yang tidak terdeteksi. Secara keseluruhan, kombinasi metrik yang tinggi, kurva ROC-AUC dan PR-AUC yang mendekati nilai maksimum, serta tidak adanya kesalahan pada klasifikasi pasien sehat

mengonfirmasi bahwa model memiliki kinerja yang sangat kuat dan dapat diandalkan dalam mendeteksi pneumonia dari citra X-ray.



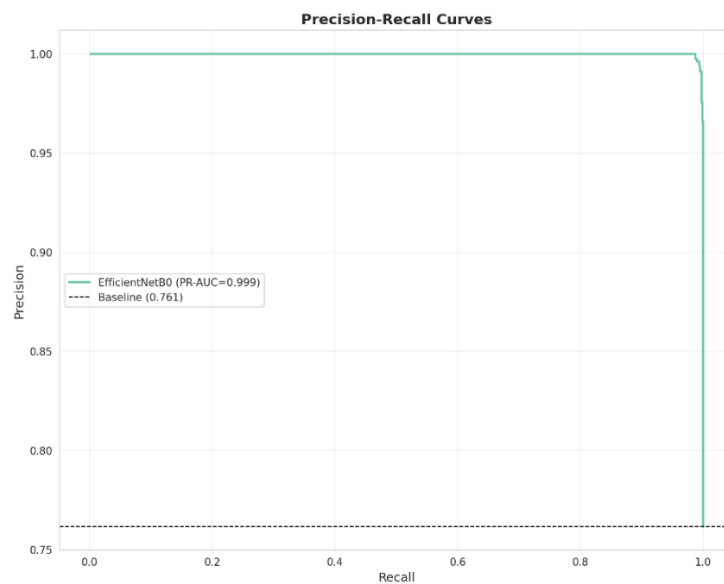
Gambar 4. 8 ROC Curve EfficientNetB0 (Skenario 1A-F)

Kurva ROC menunjukkan $AUC = 0,9996$, mendekati nilai sempurna. Kurva hampir menempel di sudut kiri atas, menandakan performa klasifikasi yang sangat baik dalam membedakan pneumonia dan normal.



Gambar 4. 9 Confusion Matrix EfficientNetB0 (Skenario 1A-F)

Confusion matrix memperlihatkan **tanpa kesalahan pada pasien normal** (TN=249, FP=0), sementara masih ada 32 kasus pneumonia yang tidak terdeteksi (FN=32). Hal ini menegaskan spesifisitas 100% dan recall 95,97%.



Gambar 4. 10 Precision-Recall Curve EfficientNetB0 (Skenario 1A-F)

Kurva Precision-Recall dengan PR-AUC = **0,9985** menunjukkan bahwa model sangat stabil meskipun distribusi kelas tidak seimbang (25,7% normal vs 74,3% pneumonia).

Interpretasi Skenario 1B-F (Split 70:30, Full Dataset)

Dari log hasil training skenario 1B-F (EfficientNetB0, 70:30, K=5), tabel hasil tiap fold dapat dilihat pada Tabel 4.7 berikut:

Tabel 4. 7 Log Training Skenario 1B-F Fold = 5

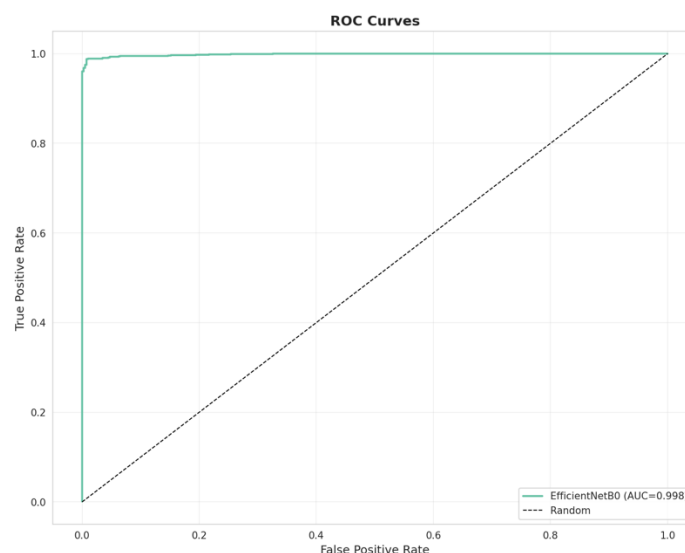
Fold	Val Accuracy
1	96,16%
2	97,26%
3	97,67%
4	97,94%
5	98,23%
Rata-rata (OOF)	97,18%

Pada skenario pengujian dengan proporsi data uji yang lebih besar (70:30), model tetap menunjukkan performa yang sangat kuat. Akurasi mencapai 97,57%, mencerminkan konsistensi kinerja meskipun jumlah sampel pengujian meningkat. Presisi yang sangat tinggi, yaitu 99,82%, menunjukkan bahwa hampir seluruh prediksi positif yang dihasilkan model benar-benar merupakan kasus pneumonia. Sensitivitas sebesar 96,90% mengindikasikan bahwa mayoritas pasien pneumonia berhasil terdeteksi, meskipun masih terdapat 36 kasus yang terlewat.

Nilai spesifisitas 99,50% menunjukkan bahwa model mampu menekan jumlah false positive secara efektif, sehingga pasien sehat hampir tidak pernah salah terklasifikasi sebagai penderita pneumonia. F1-score sebesar 98,34% menunjukkan keseimbangan yang sangat baik antara presisi dan recall. Selain itu, nilai ROC-AUC sebesar 99,84% dan PR-AUC mencapai 99,86% menegaskan bahwa model memiliki kemampuan pemisahan kelas yang sangat stabil pada berbagai threshold,

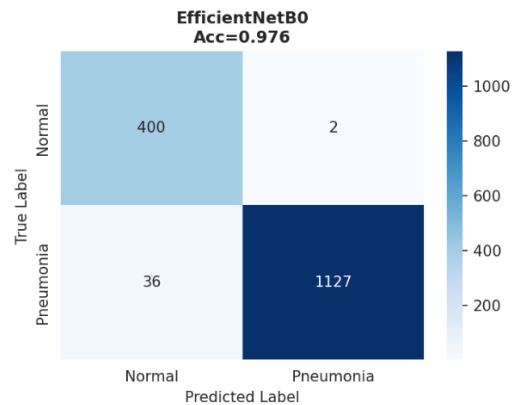
bahkan ketika diuji dengan dataset yang lebih besar.

Confusion matrix memperlihatkan total 400 true negative dan hanya 2 false positive, yang menunjukkan ketelitian model dalam mengenali pasien normal. Di sisi lain, terdapat 1127 true positive dan 36 false negative, yang menunjukkan bahwa meskipun performanya tinggi, masih ada beberapa kasus pneumonia yang tidak teridentifikasi. Secara keseluruhan, hasil evaluasi ini menegaskan bahwa model tetap andal dan efektif dalam mendeteksi pneumonia pada konfigurasi data uji yang lebih besar.



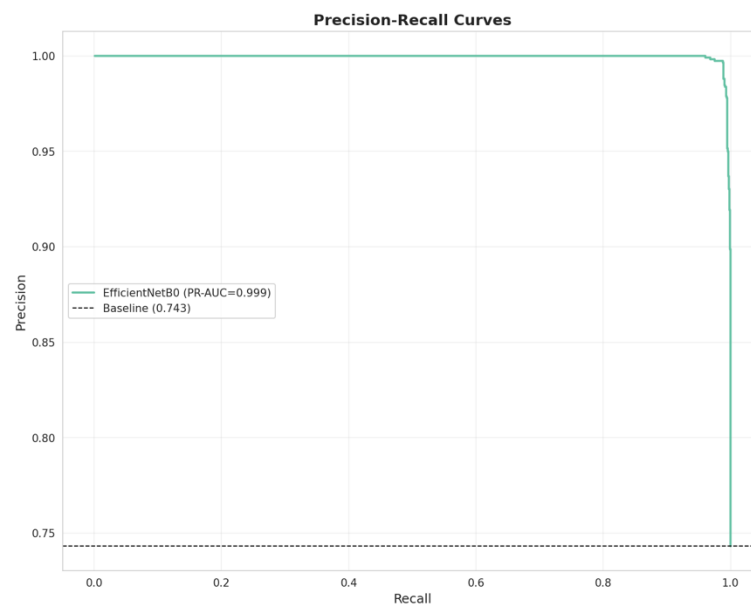
Gambar 4. 11 ROC Curve EfficientNetB0 (Skenario 1B-F)

Kurva ROC dengan $AUC = 0,9984$ memperlihatkan bahwa performa klasifikasi tetap sangat tinggi meski data uji lebih besar (70:30).



Gambar 4. 12 Confusion Matrix EfficientNetB0 (Skenario 1B-F)

Confusion matrix menunjukkan **400 pasien normal terdeteksi dengan benar**, hanya 2 salah prediksi (FP). Dari sisi pneumonia, 1127 terdeteksi benar, dan 36 luput (FN).



Gambar 4. 13 Precision-Recall Curve EfficientNetB0 (Skenario 1B-F)

Dengan PR-AUC = **0,9986**, model menunjukkan kestabilan prediksi positif yang sangat baik, bahkan pada data uji yang lebih besar.

4.2.2 Hasil Pengujian Model Tunggal: MobileNetV2

MobileNetV2 merupakan arsitektur CNN yang dirancang ringan dengan memanfaatkan depthwise separable convolution dan inverted residuals. Model ini populer untuk deployment pada perangkat terbatas sumber daya, seperti smartphone. Dalam penelitian ini, MobileNetV2 diuji pada empat skenario, yaitu 80:20 (2A-K) , 70:30 (2B-K), 80:20 (2A-F) , dan 70:30 (2B-F) menggunakan validasi silang K-Fold (K = 5).

Tabel 4. 8 Hasil Evaluasi Model MobileNetV2

Kode	Split	Akurasi	Presisi	Sensitivitas	Spesifisitas	F1-Score	AUC-ROC
2A-K	80:20	94,50%	98,51%	91,28%	98,35%	94,76%	98,63%
2B-K	70:30	95,00%	97,35%	93,04%	97,18%	95,15%	98,86%
2A-F	80:20	94,16%	99,20%	93,08%	97,59%	96,04%	98,98%
2B-F	70:30	91,31%	99,42%	88,82%	98,51%	93,82%	98,98%

Interpretasi Skenario 2A-K (Split 80:20)

Dari log hasil training skenario **2A-K** (MobileNetV2, 80:20, K=5), tabel hasil tiap fold bisa dilihat pada table 4.9 berikut:

Tabel 4. 9 Log Training Skenario 2A-K Fold = 5

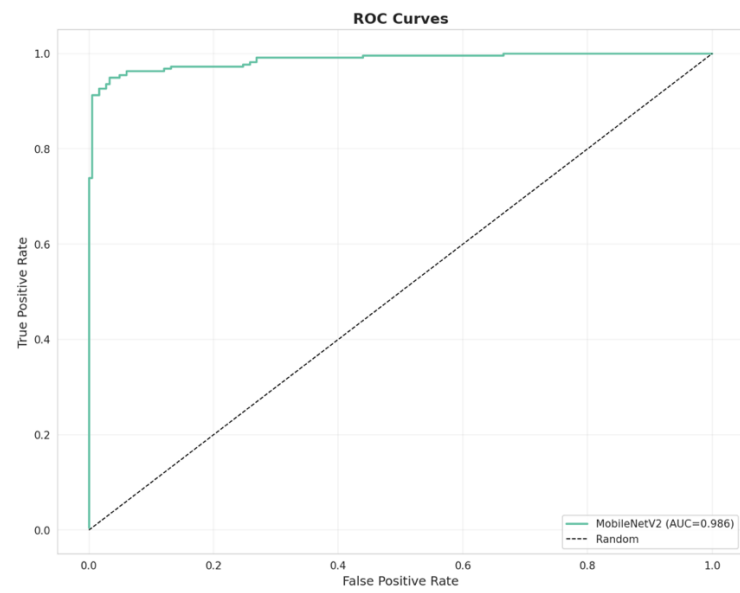
Fold	Val Accuracy
1	71,47%
2	88,71%
3	64,89%
4	75,55%
5	86,11%
Rata-rata (OOF)	86,50%

Hasil pengujian pada model MobileNetV2 menunjukkan bahwa performanya berada sedikit di bawah EfficientNetB0, namun tetap berada pada kategori tinggi. Akurasi yang diperoleh sebesar 94,50%, yang masih menunjukkan kemampuan klasifikasi yang kuat. Presisi sebesar 98,51% dan spesifisitas 98,35% mengindikasikan bahwa model sangat jarang melakukan kesalahan dalam

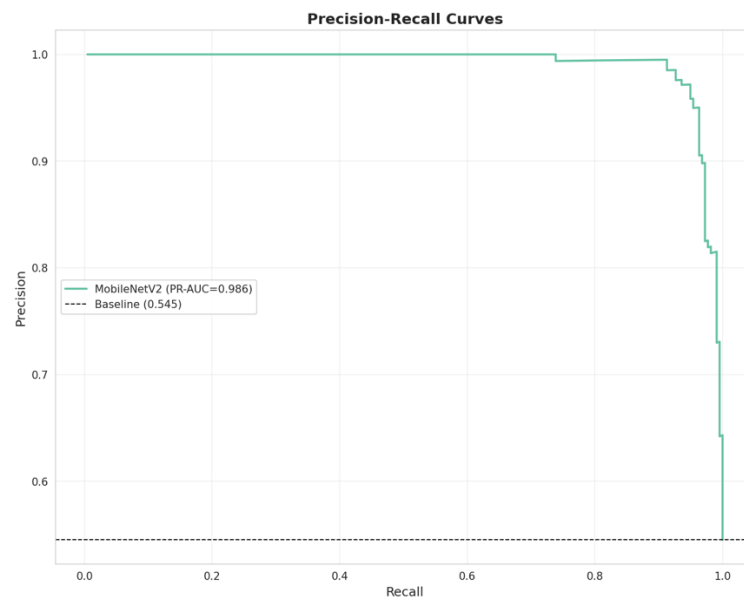
mengklasifikasikan pasien normal, sehingga jumlah false positive relatif rendah.

Sensitivitas mencapai 91,28%, lebih rendah dibandingkan model EfficientNetB0, yang menunjukkan masih terdapat sejumlah kasus pneumonia yang tidak terdeteksi. Hal ini juga terlihat pada confusion matrix, di mana dari total 400 sampel, model menghasilkan 199 true positive, 179 true negative, hanya 3 false positive, namun terdapat 19 false negative. Kondisi ini memperlihatkan bahwa meskipun model cukup akurat, tingkat ketidakberhasilan dalam mendeteksi pneumonia masih lebih tinggi dibandingkan model lainnya.

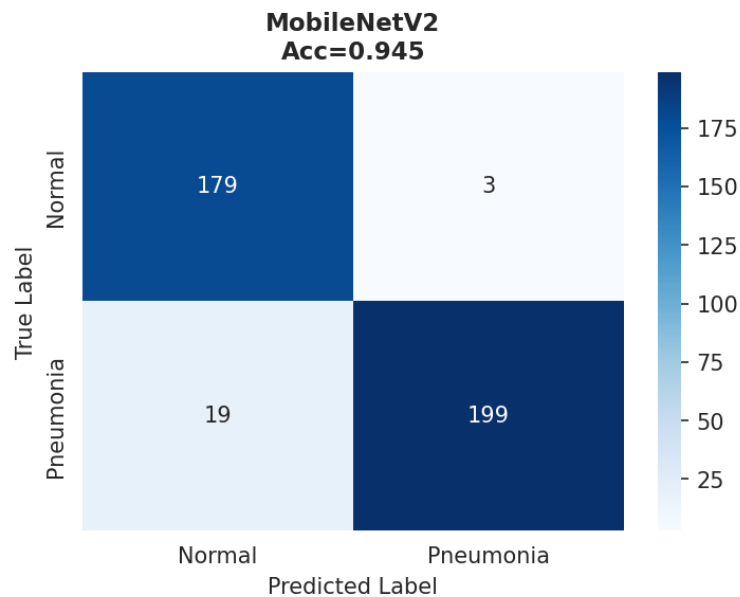
Evaluasi kurva ROC yang ditampilkan pada Gambar 4.14 menunjukkan nilai AUC sebesar 0,9863, sedikit lebih rendah dibandingkan EfficientNetB0, tetapi masih tergolong sangat baik dan menandakan kemampuan pemisahan kelas yang kuat. Selain itu, Precision-Recall Curve pada Gambar 4.15 menghasilkan nilai PR-AUC sebesar 0,9861 yang menegaskan bahwa performa model tetap stabil meskipun sensitivitas tidak setinggi model terbaik. Secara keseluruhan, model MobileNetV2 menunjukkan performa kompetitif dengan tingkat kesalahan yang rendah, terutama pada klasifikasi pasien normal.



Gambar 4. 14 ROC Curve MobileNetV2 (Skenario 2A-K)



Gambar 4. 15 Precision-Recall Curve MobileNetV2 (Skenario 2A-K)



Gambar 4. 16 Confusion Matrix MobileNetV2 (Skenario 2A-K)

Interpretasi Skenario 2B-K (Split 70:30)

Dari log hasil training skenario **2B-K** (MobileNetV2, 70:30, K=5), tabel hasil tiap fold bisa dilihat pada table 4.10 berikut:

Tabel 4. 10 Log Training Skenario 2B-K Fold = 5

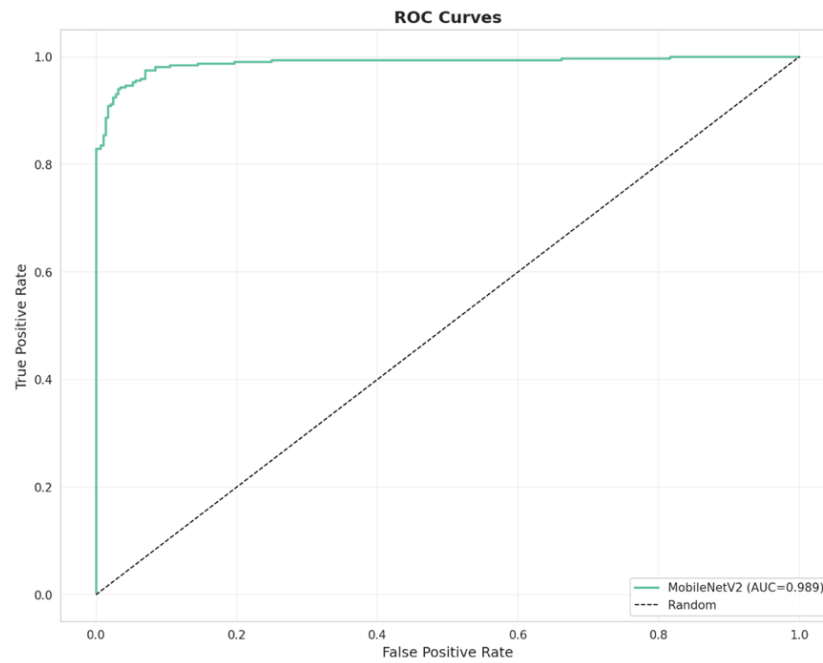
Fold	Val Accuracy
1	71,47%
2	88,71%
3	64,89%
4	75,55%
5	86,11%
Rata-rata (OOF)	86,50%

Pada skenario 2B-K, model MobileNetV2 menunjukkan peningkatan performa dibandingkan skenario sebelumnya. Akurasi yang diperoleh mencapai 95,00%, sedikit lebih tinggi daripada skenario 2A-K, menandakan peningkatan stabilitas model meskipun proporsi data latih lebih kecil. Presisi sebesar 97,35% dan spesifisitas 97,18% tetap berada pada tingkat yang tinggi, sehingga model konsisten dalam menekan jumlah false positive dan menjaga ketepatan klasifikasi

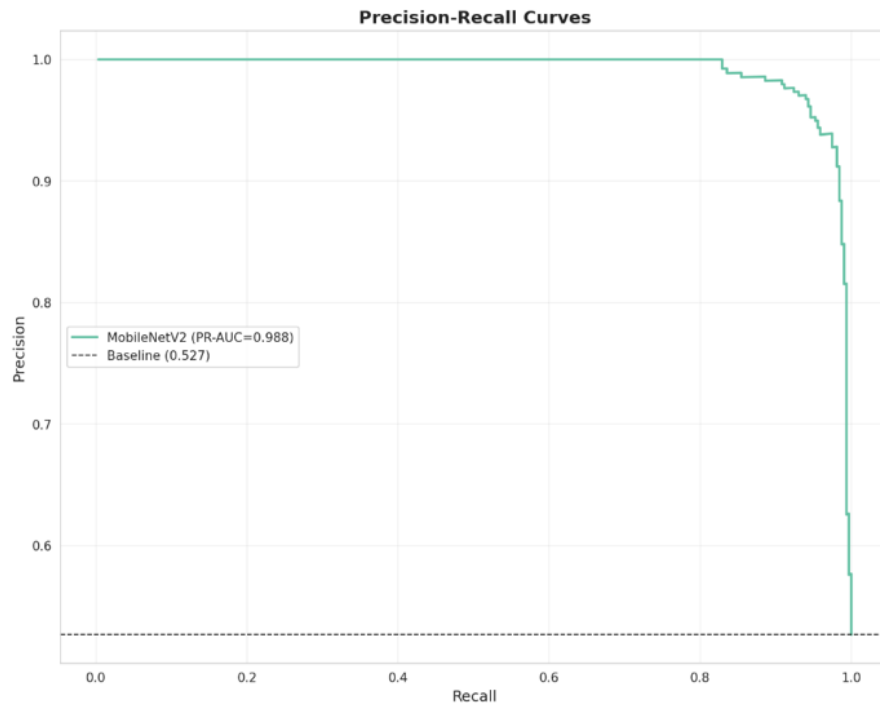
terhadap pasien normal.

Sensitivitas meningkat menjadi 93,04%, menunjukkan bahwa model mampu mendeteksi lebih banyak kasus pneumonia dibandingkan skenario sebelumnya. Hal ini tercermin pada confusion matrix, di mana dari total 600 sampel, model menghasilkan 294 true positive dan 276 true negative, disertai 8 false positive dan 22 false negative. Distribusi ini mengindikasikan peningkatan kemampuan model dalam menangkap kasus positif tanpa mengorbankan ketelitian pada kasus negatif.

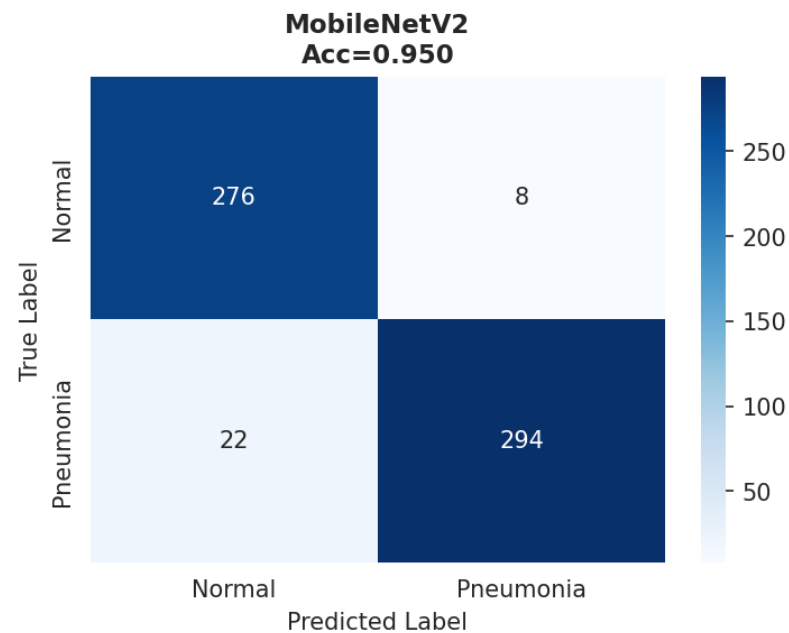
Kurva ROC pada Gambar 4.16 memperlihatkan nilai AUC sebesar 0,9886, yang sedikit lebih tinggi dibandingkan skenario 2A-K, menandakan pemisahan kelas yang lebih baik. Precision-Recall Curve pada Gambar 4.17 menunjukkan nilai PR-AUC sebesar 0,9884, sehingga performa model dinilai tetap stabil di berbagai threshold. Secara keseluruhan, skenario 2B-K memperlihatkan bahwa model MobileNetV2 dapat mempertahankan dan bahkan sedikit meningkatkan performa meskipun konfigurasi data latih berubah.



Gambar 4. 17 ROC Curve MobileNetV2 (Skenario 2B-K)



Gambar 4. 18 Precision-Recall Curve MobileNetV2 (Skenario 2B-K)



Gambar 4. 19 Confusion Matrix MobileNetV2 (Skenario 2B-K)

Interpretasi Skenario 2A-F (Split 80:20, Full Dataset)

Dari log hasil training skenario 2A-F (MobileNetV2, 80:20, K=5), tabel hasil tiap fold dapat dilihat pada **Tabel 4.11** berikut:

Tabel 4. 11 Log Training Skenario 2A-F Fold = 5

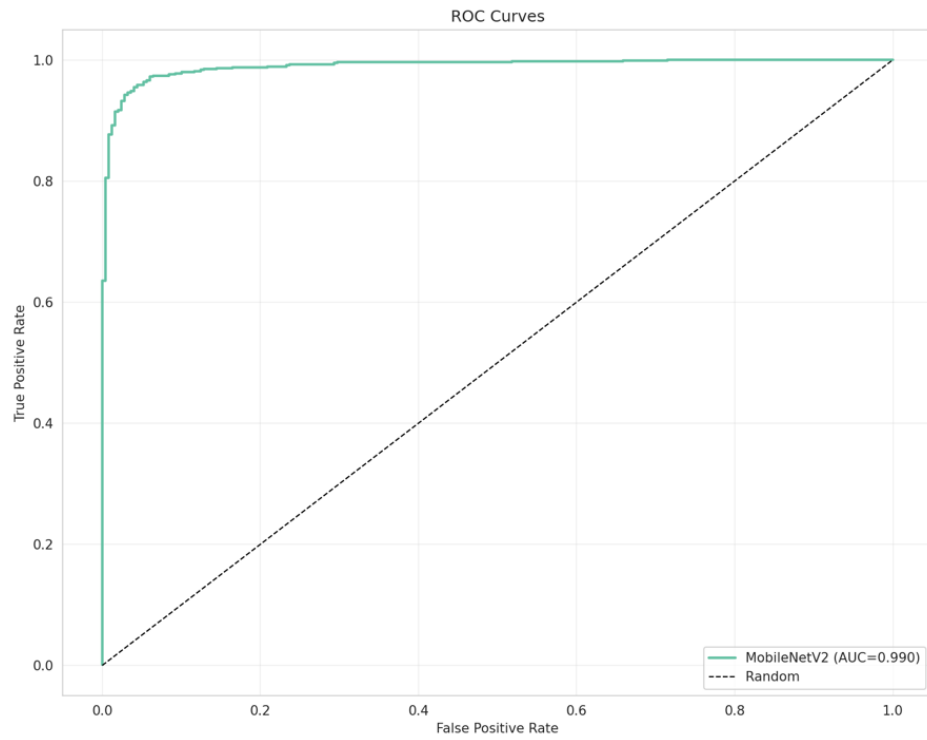
Fold	Val Accuracy
1	92,93%
2	83,69%
3	87,05%
4	83,69%
5	87,20%
Rata-rata (OOF)	90,77%

Pada skenario 2A-F, model MobileNetV2 menunjukkan performa yang sangat baik ketika diuji pada keseluruhan dataset. Akurasi yang dicapai adalah 94,16%, yang menunjukkan kemampuan klasifikasi yang kuat pada data berukuran besar. Nilai presisi yang tinggi, yaitu 99,20%, serta spesifisitas sebesar 97,59%, mengindikasikan bahwa model sangat jarang melakukan kesalahan dalam mengklasifikasikan pasien normal, sehingga jumlah false positive tetap rendah.

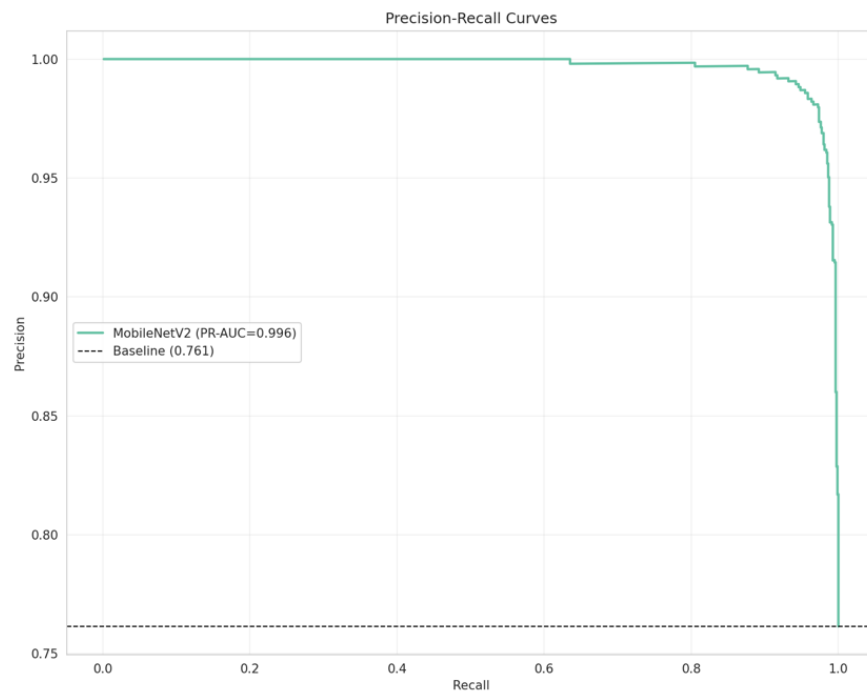
Meskipun demikian, nilai sensitivitas sebesar 93,08% masih lebih rendah dibandingkan model EfficientNetB0, yang berarti masih terdapat beberapa kasus pneumonia yang tidak terdeteksi dengan baik.

Confusion matrix mendukung temuan tersebut, dengan hasil 243 true negative dan hanya 6 false positive, namun terdapat 55 false negative dari total 740 pasien pneumonia yang benar. Hal ini menunjukkan bahwa meskipun model sangat baik dalam menjaga presisi dan spesifisitas, kemampuan menangkap seluruh kasus positif masih dapat ditingkatkan.

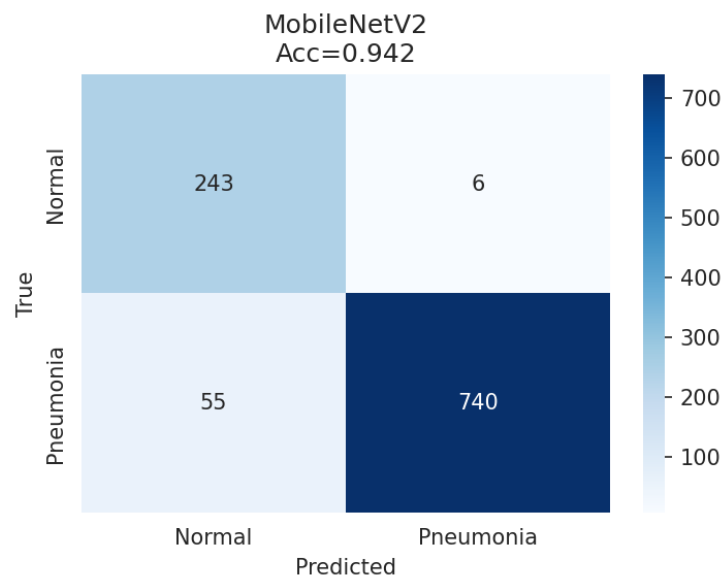
Kurva ROC pada Gambar 4.19 memperlihatkan nilai AUC sebesar 0,9898, yang menandakan kemampuan pemisahan kelas yang sangat baik. Precision–Recall Curve pada Gambar 4.20 menghasilkan PR-AUC sebesar 0,9955, menunjukkan bahwa model tetap stabil dan akurat meskipun terdapat ketidakseimbangan antara jumlah kelas positif dan negatif. Secara keseluruhan, performa pada skenario 2A-F menegaskan bahwa MobileNetV2 tetap mampu memberikan hasil yang kuat dan dapat diandalkan pada dataset penuh.



Gambar 4. 20 ROC Curve MobileNetV2 (Skenario 2A-F)



Gambar 4. 21 Precision–Recall Curve MobileNetV2 (Skenario 2A-F)



Gambar 4. 22 Confusion Matrix MobileNetV2 (Skenario 2A-F)

Interpretasi Skenario 2B-F (Split 70:30, Full Dataset)

Dari log hasil training skenario 2B-F (MobileNetV2, 70:30, K=5), tabel hasil tiap fold dapat dilihat pada **Tabel 4.12** berikut:

Tabel 4. 12 Log Training Skenario 2B-F Fold = 5

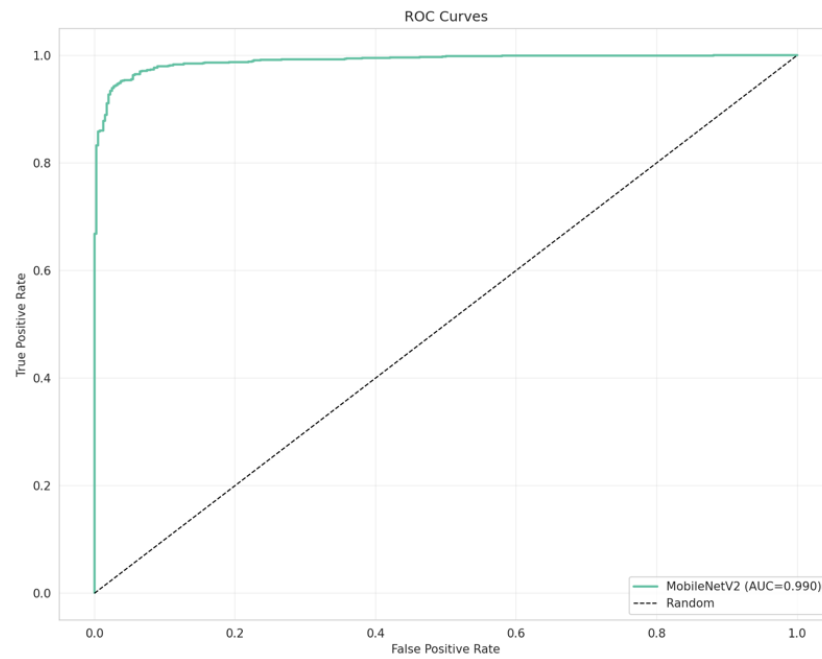
Fold	Val Accuracy
1	79,15%
2	82,17%
3	85,87%
4	85,19%
5	94,01%
Rata-rata (OOF)	86,58%

Pada skenario 2B-F, model MobileNetV2 tetap menunjukkan performa yang kuat meskipun jumlah data uji diperbesar. Akurasi mencapai 91,31%, menandakan bahwa model masih mampu melakukan klasifikasi dengan baik pada kondisi pengujian yang lebih menantang. Presisi yang sangat tinggi, yaitu 99,42%, menunjukkan bahwa hampir seluruh prediksi positif yang dihasilkan model benar-benar merupakan kasus pneumonia. Namun, sensitivitas mengalami penurunan

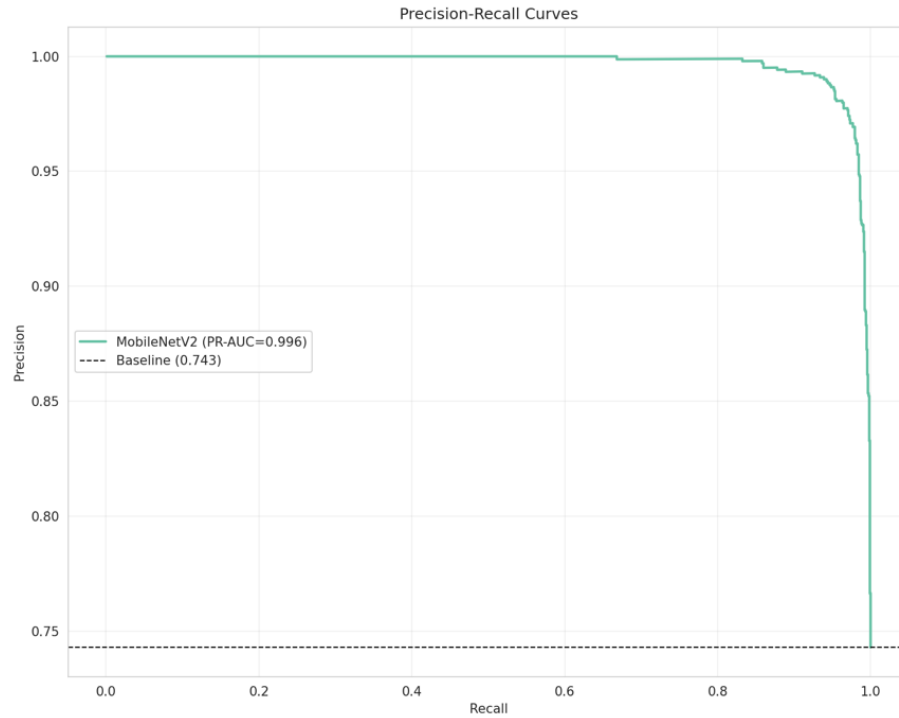
menjadi 88,82%, yang mengindikasikan bahwa semakin banyak kasus pneumonia yang terlewat ketika proporsi data uji bertambah besar.

Confusion matrix memperlihatkan bahwa model menghasilkan 396 true negative dan hanya 6 false positive, sehingga kemampuan model dalam mengenali pasien sehat tetap stabil dengan spesifisitas 98,51%. Di sisi lain, terdapat 1033 true positive dan 130 false negative, yang mencerminkan peningkatan jumlah kasus pneumonia yang tidak terdeteksi dibandingkan skenario sebelumnya. Meskipun demikian, nilai F1-score sebesar 93,82% menunjukkan bahwa keseimbangan antara presisi dan recall masih cukup baik.

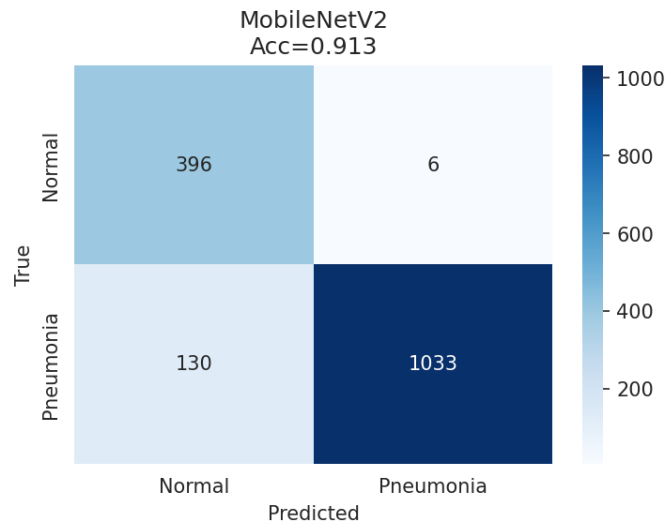
Kurva ROC yang ditampilkan pada Gambar 4.22 menghasilkan AUC sebesar 0,9898, menandakan bahwa model tetap memiliki kemampuan pemisahan kelas yang sangat baik di berbagai threshold. Demikian pula, Precision–Recall Curve pada Gambar 4.23 memberikan nilai PR-AUC sebesar 0,9957, yang menunjukkan bahwa performa klasifikasi tetap kuat meskipun terjadi penurunan sensitivitas. Secara keseluruhan, hasil pada skenario 2B-F menegaskan bahwa model tetap stabil dan efektif meskipun dihadapkan pada proporsi data uji yang lebih besar.



Gambar 4. 23 ROC Curve MobileNetV2 (Skenario 2B-F)



Gambar 4. 24 Precision-Recall Curve MobileNetV2 (Skenario 2B-F)



Gambar 4. 25 Confusion Matrix MobileNetV2 (Skenario 2B-F)

4.2.3 Hasil Pengujian Model Tunggal: MobileNetV3Small

MobileNetV3Small merupakan varian dari MobileNetV3 yang lebih ringan, dengan optimasi melalui *NAS (Neural Architecture Search)* dan penggunaan blok *squeeze-and-excitation*. Arsitektur ini didesain untuk efisiensi tinggi pada perangkat mobile, namun kapasitas representasinya lebih kecil dibanding EfficientNetB0 maupun MobileNetV2.

Dalam penelitian ini, MobileNetV3Small diuji pada empat skenario , yaitu 80:20 (3A-K) , 70:30 (3B-K), 80:20 (3A-F), dan 70:30 (3B-F) menggunakan validasi silang K-Fold (K = 5).

Tabel 4. 13 Hasil Evaluasi Model MobileNetV3Small

Kode	Split	Akurasi	Presisi	Sensitivitas	Spesifisitas	F1-Score	AUC-ROC
3A-K	80:20	92,75%	100,00%	86,70%	100,00%	92,87%	99,44%
3B-K	70:30	71,33%	100,00%	45,57%	100,00%	62,61%	98,46%
3A-F	80:20	95,31%	99,08%	94,72%	97,19%	96,85%	99,41%
3B-F	70:30	91,25%	99,14%	88,99%	97,76%	93,79%	98,88%

Interpretasi Skenario 3A-K (Split 80:20)

Dari log hasil training skenario 3A-K (MobileNetV3Small, 80:20, K=5), tabel hasil tiap fold dapat dilihat pada **Tabel 4.14** berikut:

Tabel 4. 14 Log Training Skenario 3A-K Fold = 5

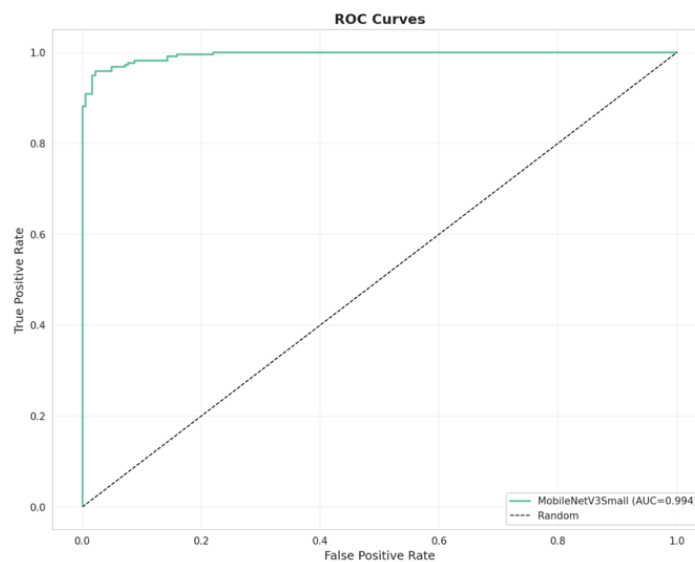
Fold	Val Accuracy
1	57,37%
2	58,62%
3	87,15%
4	84,95%
5	85,80%
Rata-rata (OOF)	75,69%

Pada skenario pengujian ini, model MobileNetV3Small menunjukkan performa yang relatif lebih rendah dibandingkan EfficientNetB0 dan MobileNetV2. Akurasi yang dicapai sebesar 92,75%, yang meskipun masih tergolong tinggi, tetap berada di bawah dua model sebelumnya. Nilai presisi dan spesifisitas masing-masing mencapai 100%, menunjukkan bahwa model tidak melakukan kesalahan dalam mengklasifikasikan pasien normal dan tidak menghasilkan false positive sama sekali. Hal ini menandakan ketelitian yang sangat kuat dalam mendeteksi kelas negatif.

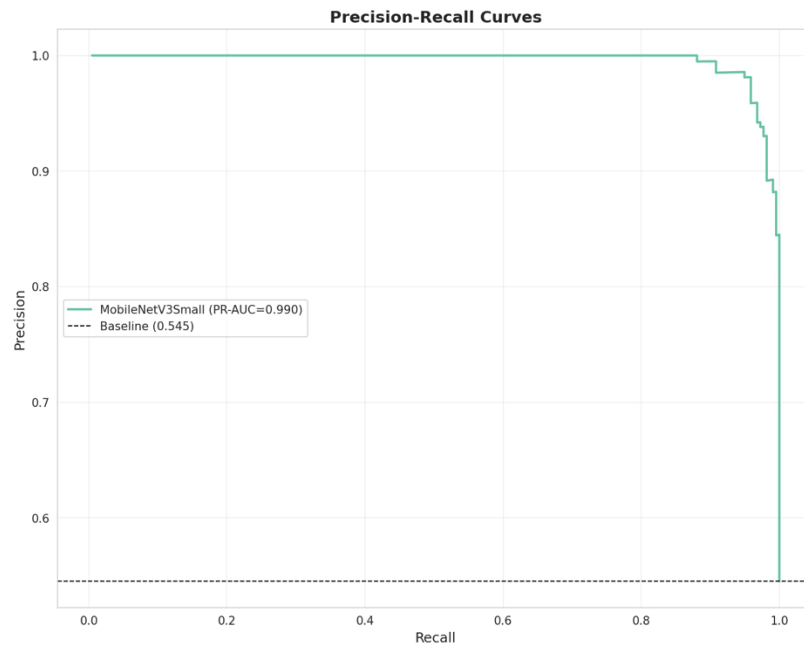
Namun, sensitivitas model hanya mencapai 86,70%, jauh lebih rendah dibandingkan nilai sensitivitas EfficientNetB0 yang mencapai 98,62%. Kondisi ini mengindikasikan bahwa masih terdapat cukup banyak kasus pneumonia yang tidak

terdeteksi, sebagaimana terlihat pada confusion matrix dengan 29 false negative dari total 400 sampel. Secara keseluruhan, model menghasilkan 189 true positive dan 182 true negative, sementara tidak terdapat false positive pada skenario ini.

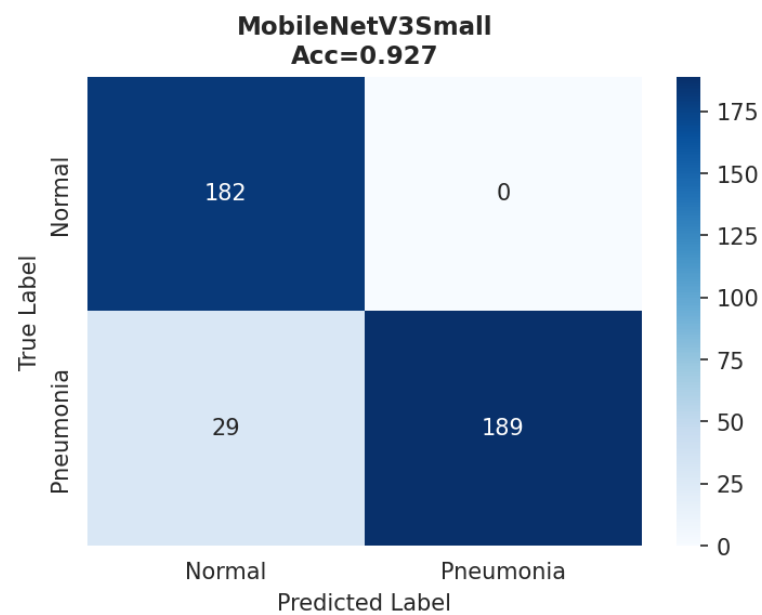
Evaluasi melalui kurva ROC pada Gambar 4.25 menunjukkan nilai AUC sebesar 0,9944, yang tetap menandakan kemampuan pemisahan kelas yang sangat baik meskipun sensitivitas lebih rendah. Precision–Recall Curve pada Gambar 4.26 memberikan nilai PR-AUC sebesar 0,9903. Meskipun nilainya tinggi, bentuk kurva terlihat lebih curam akibat penurunan recall. Secara keseluruhan, hasil ini menunjukkan bahwa MobileNetV3Small memiliki presisi dan spesifisitas yang sangat kuat, tetapi performa deteksi kasus positif masih tertinggal dibandingkan model lain.



Gambar 4. 26 ROC Curve MobileNetV3Small (Skenario 3A-K)



Gambar 4. 27 Precision-Recall Curve MobileNetV3Small (Skenario 3A-K)



Gambar 4. 28 Confusion Matrix MobileNetV3Small (Skenario 3A-K)

Interpretasi Skenario 3B-K (Split 70:30)

Dari log hasil training skenario 3B-K (MobileNetV3Small, 70:30, K=5), tabel hasil tiap fold dapat dilihat pada **Tabel 4.15** berikut:

Tabel 4. 15 Log Training Skenario 3B-K Fold = 5

Fold	Val Accuracy
1	83,15%
2	73,12%
3	48,75%
4	72,40%
5	73,59%
Rata-rata (OOF)	70,86%

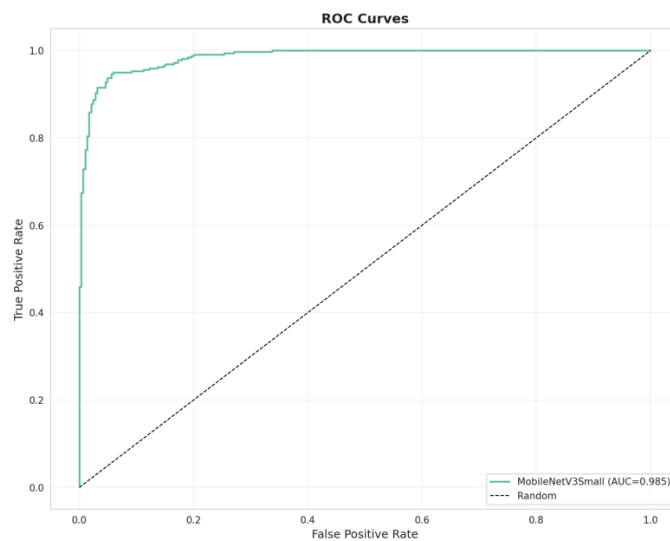
Pada skenario ini, performa model MobileNetV3Small mengalami penurunan yang sangat signifikan ketika proporsi data uji diperbesar. Akurasi turun drastis menjadi 71,33%, menunjukkan bahwa model tidak stabil dan kesulitan mempertahankan kualitas prediksi pada konfigurasi data yang lebih besar. Meskipun demikian, nilai presisi dan spesifisitas tetap berada pada angka 100%, yang menunjukkan bahwa model tidak pernah melakukan kesalahan dalam mengidentifikasi pasien normal, sehingga tidak terdapat false positive.

Kontras dengan hal tersebut, sensitivitas anjlok ke 45,57%, menandakan bahwa lebih dari separuh pasien pneumonia tidak berhasil terdeteksi. Hal ini terlihat jelas pada confusion matrix, di mana dari total 600 sampel, model menghasilkan 144 true positive dan 284 true negative, tetapi juga mencatat 172 false negative. Tingginya jumlah kasus yang terlewat ini merupakan kelemahan kritis, terutama pada konteks diagnosis medis yang membutuhkan deteksi positif yang optimal.

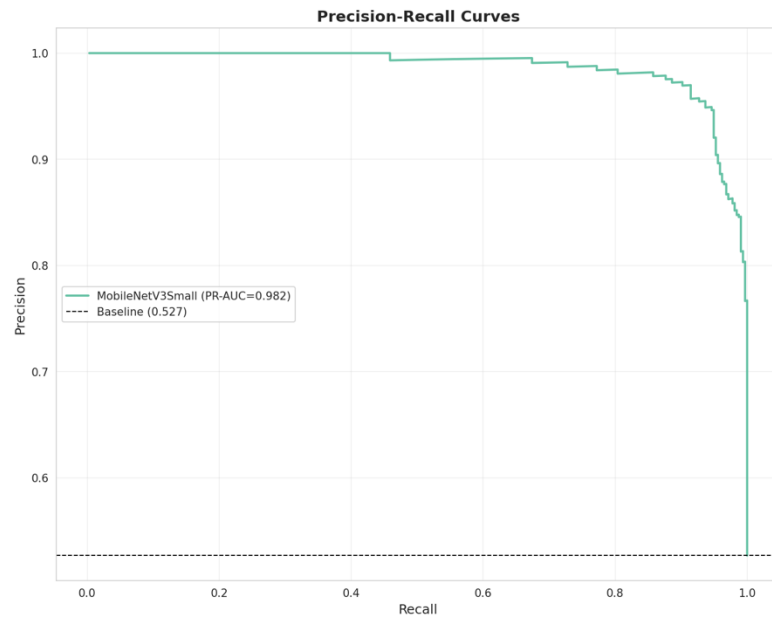
Kurva ROC yang ditunjukkan pada Gambar 4.28 masih menghasilkan nilai AUC sebesar 0,9846, yang sebenarnya menandakan bahwa model mampu menghasilkan probabilitas yang cukup baik dalam memisahkan kelas. Namun,

performa aktual pada threshold default 0,5 justru buruk karena banyak prediksi positif yang gagal terdeteksi. Precision–Recall Curve pada Gambar 4.29 memberikan nilai PR-AUC sebesar 0,9825. Meskipun nilainya tinggi, kurva tersebut menunjukkan adanya kesenjangan besar antara presisi yang sempurna dan recall yang sangat rendah.

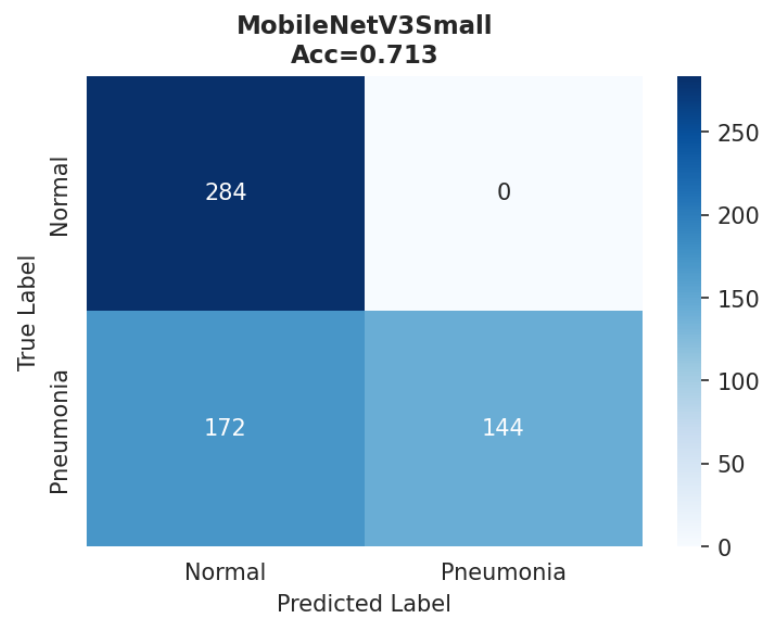
Secara keseluruhan, hasil pada skenario ini menunjukkan bahwa MobileNetV3Small tidak mampu mempertahankan stabilitas performa ketika proporsi data uji diperbesar, terutama dalam mendeteksi kasus pneumonia secara konsisten.



Gambar 4. 29 ROC Curve MobileNetV3Small (Skenario 3B-K)



Gambar 4. 30 Precision-Recall Curve MobileNetV3Small (Skenario 3B-K)



Gambar 4. 31 Confusion Matrix MobileNetV3Small (Skenario 3B-K)

Interpretasi Skenario 3A-F (Split 80:20, Full Dataset)

Dari log hasil training skenario 3A-F (MobileNetV3Small, 80:20, K=5), tabel hasil tiap fold dapat dilihat pada **Tabel 4.16** berikut:

Tabel 4. 16 Log Training Skenario 3A-F Fold = 5

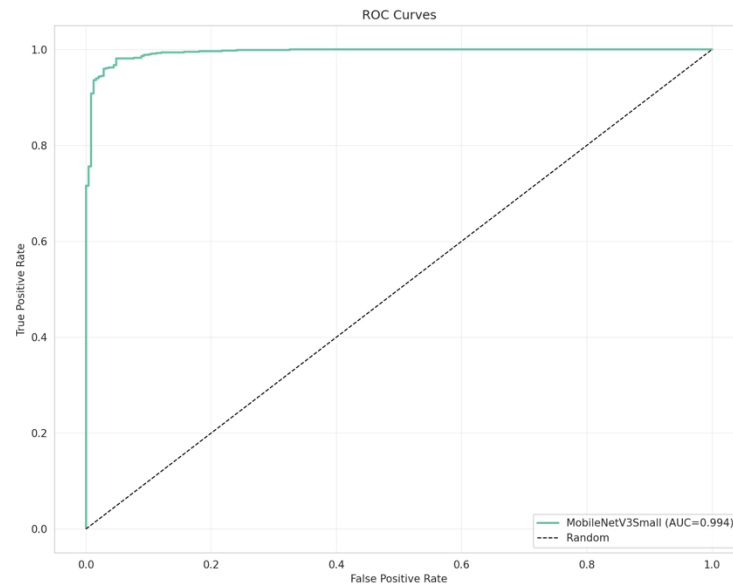
Fold	Val Accuracy
1	82,73%
2	79,38%
3	77,58%
4	85,49%
5	74,88%
Rata-rata (OOF)	85,86%

Pada skenario ini, performa model MobileNetV3Small menunjukkan peningkatan yang signifikan dibandingkan hasil pada K-Fold sebelumnya (3A-K). Akurasi mencapai 95,31%, yang menandakan bahwa pelatihan dengan dataset penuh mampu meningkatkan stabilitas dan ketepatan prediksi model. Nilai presisi sebesar 99,08% dan spesifisitas 97,19% menunjukkan bahwa model sangat jarang melakukan kesalahan dalam mendeteksi pasien normal, sehingga jumlah false positive tetap sangat rendah.

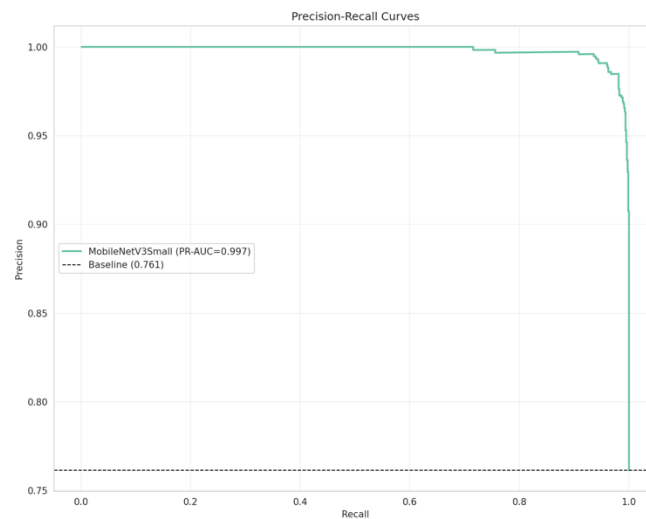
Sensitivitas meningkat tajam menjadi 94,72%, jauh lebih tinggi dibandingkan nilai 86,70% pada skenario 3A-K. Peningkatan ini mengindikasikan bahwa model menjadi lebih efektif dalam mengenali kasus pneumonia setelah mendapatkan lebih banyak variasi data selama proses pelatihan. Hal tersebut juga tercermin pada confusion matrix, di mana dari total 1.044 sampel, model menghasilkan 753 true positive dan 242 true negative, disertai 7 false positive dan 42 false negative.

Evaluasi melalui kurva ROC yang ditampilkan pada Gambar 4.31 menghasilkan nilai AUC sebesar 0,9941, yang menunjukkan kemampuan separasi antar kelas yang sangat kuat. Precision–Recall Curve pada Gambar 4.32 memberikan nilai PR-AUC sebesar 0,9968, menandakan bahwa performa model tetap stabil meskipun terdapat ketidakseimbangan rasio antara kelas positif dan

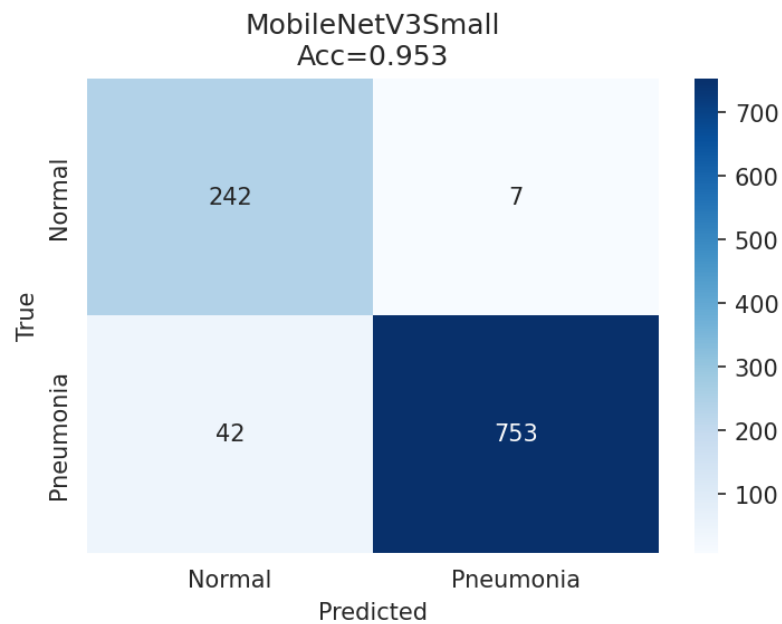
negatif. Secara keseluruhan, hasil ini menunjukkan bahwa penggunaan dataset penuh secara efektif meningkatkan kemampuan model dalam mendeteksi pneumonia tanpa mengorbankan ketelitian terhadap kelas normal.



Gambar 4. 32 ROC Curve MobileNetV3Small (Skenario 3A-F)



Gambar 4. 33 Precision–Recall Curve MobileNetV3Small (Skenario 3A-F)



Gambar 4. 34 Confusion Matrix MobileNetV3Small (Skenario 3A-F)

Interpretasi Skenario 3B-F (Split 70:30, Full Dataset)

Dari log hasil training skenario 3B-F (MobileNetV3Small, 70:30, K=5), tabel hasil tiap fold dapat dilihat pada **Tabel 4.18** berikut:

Tabel 4. 17 Log Training Skenario 3B-F Fold = 5

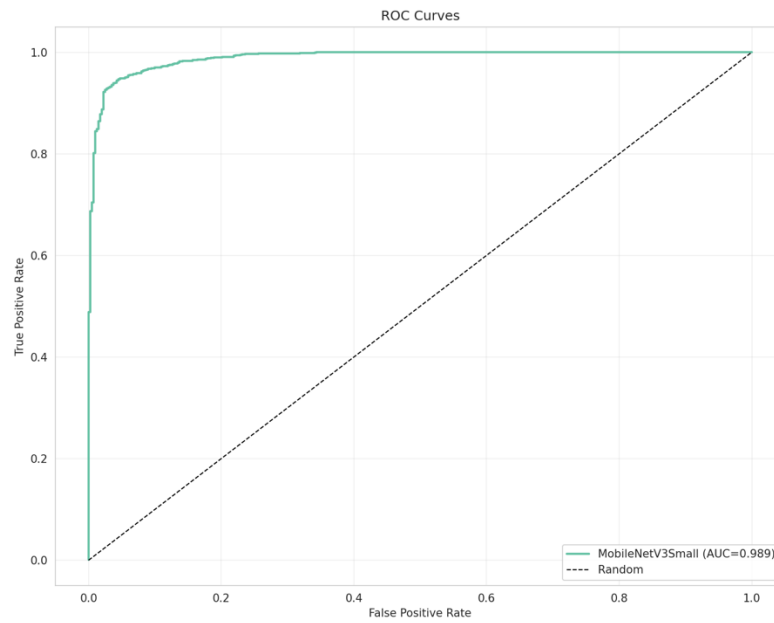
Fold	Val Accuracy
1	80,25%
2	75,17%
3	78,05%
4	74,49%
5	89,39%
Rata-rata (OOB)	83,43%

Pada skenario ini, model MobileNetV3Small menunjukkan performa yang masih kuat meskipun terjadi penurunan pada beberapa metrik akibat proporsi data uji yang lebih besar. Akurasi yang diperoleh adalah 91,25%, sedikit lebih rendah dibandingkan skenario 3A-F, namun tetap berada pada tingkat yang dapat diterima untuk tugas klasifikasi medis. Nilai presisi sebesar 99,14% dan spesifisitas 97,76%

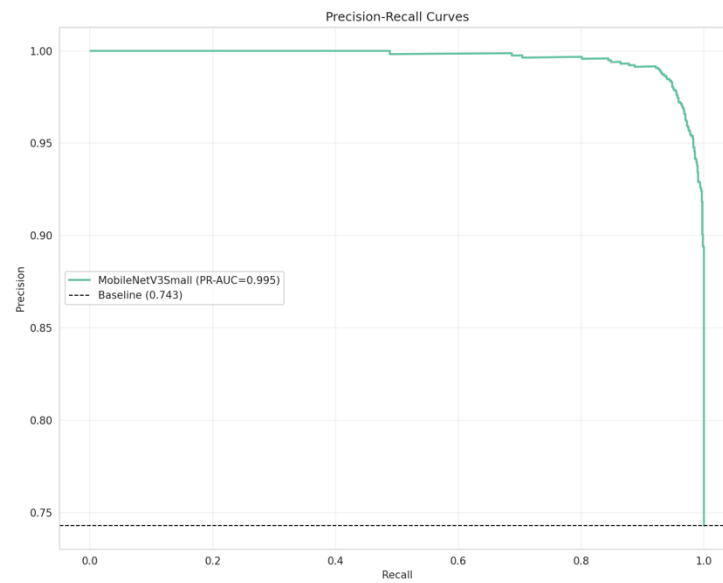
menunjukkan bahwa model tetap sangat stabil dalam menghindari kesalahan klasifikasi pada pasien normal, sehingga false positive dapat ditekan pada tingkat yang sangat rendah.

Di sisi lain, sensitivitas turun menjadi 88,99%, lebih rendah dibandingkan skenario 3A-F. Hal ini mengindikasikan bahwa sebagian kasus pneumonia masih tidak terdeteksi, sebagaimana tercermin pada confusion matrix dengan 128 false negative dari total 1.163 kasus positif. Secara keseluruhan, model menghasilkan 1.035 true positive dan 393 true negative, disertai 9 false positive dan 128 false negative dari total 1.565 sampel.

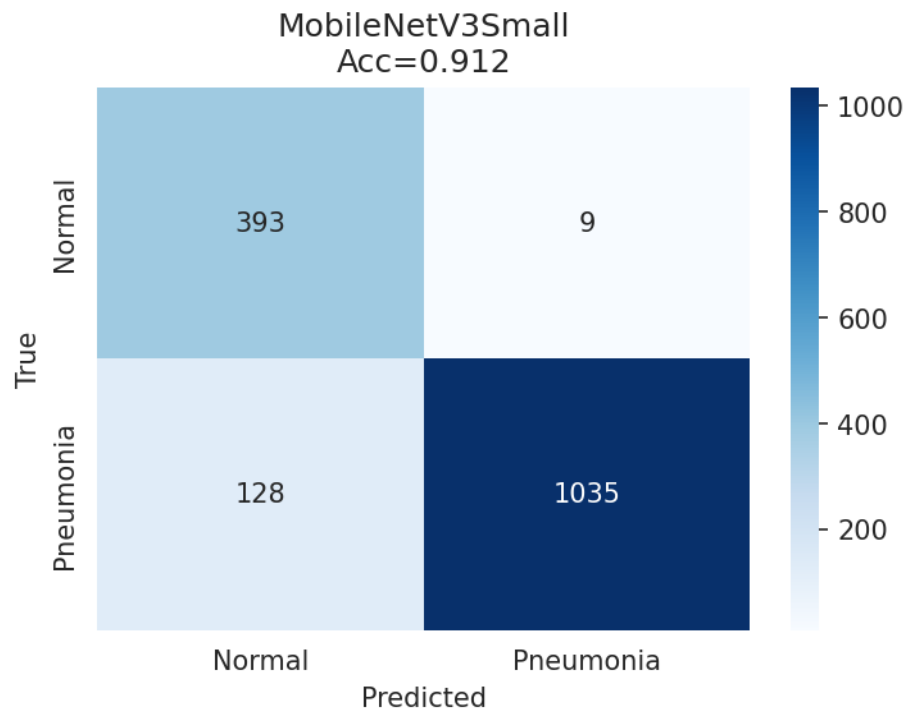
Kurva ROC pada Gambar 4.34 menunjukkan nilai AUC sebesar 0,9888, yang menandakan bahwa kemampuan diskriminatif model tetap kuat meskipun sensitivitas menurun. Precision–Recall Curve pada Gambar 4.35 menghasilkan nilai PR-AUC sebesar 0,9950, menunjukkan bahwa model masih mampu mempertahankan presisi yang sangat tinggi meskipun recall tidak setinggi skenario sebelumnya. Secara keseluruhan, hasil ini menunjukkan bahwa model masih efektif dan stabil, namun sensitif terhadap perubahan proporsi data uji terutama dalam mendeteksi kasus positif.



Gambar 4. 35 ROC Curve MobileNetV3Small (Skenario 3B-F)



Gambar 4. 36 Precision-Recall Curve MobileNetV3Small (Skenario 3B-F)



Gambar 4. 37 Confusion Matrix MobileNetV3Small (Skenario 3B-F)

4.2.4 Hasil Pengujian Model Stacking Ensemble

Stacking ensemble adalah teknik yang menggabungkan prediksi dari beberapa model dasar (base models) untuk menghasilkan prediksi akhir yang lebih robust. Pada penelitian ini, base models terdiri dari EfficientNetB0, MobileNetV2, dan MobileNetV3Small, sedangkan meta-learner yang digunakan adalah Dense Neural Network (MLP).

Pengujian dilakukan pada empat skenario, yaitu 80:20 (4A-K) ,70:30 (4B-K), 80:20 (4A-F), dan 70:30(4B-F) dengan validasi silang K-Fold ($K = 5$).

Tabel 4. 18 Hasil Evaluasi Model Stacking Ensemble

Kode	Split	Akurasi	Presisi	Sensitivitas	Spesifisitas	F1-Score	AUC-ROC
4A-K	80:20	97,25%	98,14%	96,79%	97,80%	97,46%	99,65%
4B-K	70:30	96,67%	96,84%	96,84%	96,48%	96,84%	99,70%
4A-F	80:20	99,14%	99,87%	98,99%	99,60%	99,43%	99,87%
4B-F	70:30	98,21%	99,74%	97,85%	99,25%	98,78%	99,84%

Interpretasi Skenario 4A-K (Split 80:20)

Dari log hasil training skenario 4A-K (Stacking Ensemble, 80:20, K=5), tabel hasil tiap fold dapat dilihat pada **Tabel 4.21** berikut:

Tabel 4. 19 Hasil Validasi Tiap Fold Base Models pada Skenario 4A-K

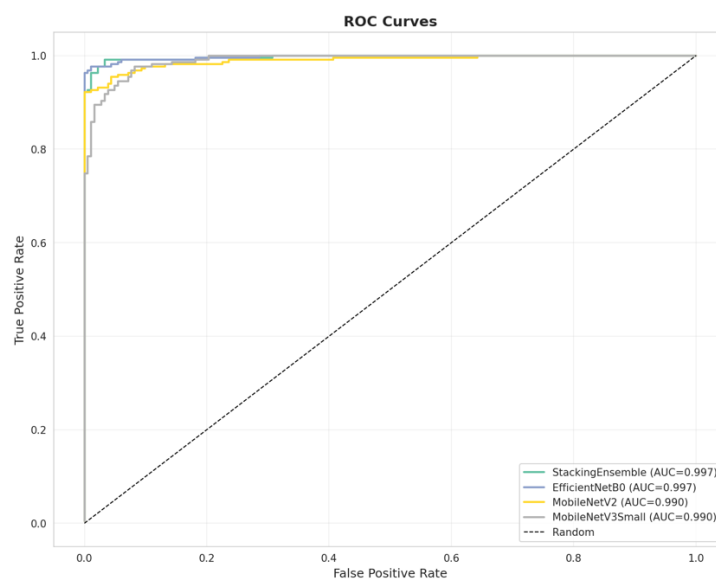
Base Model	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	OOF Accuracy
EfficientNetB0	87,77%	94,04%	91,54%	96,87%	94,44%	93,81%
MobileNetV2	90,28%	84,33%	74,61%	93,10%	88,27%	88,94%
MobileNetV3Small	63,95%	80,88%	63,32%	76,49%	90,74%	77,00%

Pada skenario ini, model stacking ensemble menunjukkan performa yang kompetitif dibandingkan model tunggal terbaik, EfficientNetB0. Akurasi yang diperoleh sebesar 97,25%, sedikit lebih rendah dari nilai 97,75% milik EfficientNetB0, namun tetap berada dalam kategori sangat tinggi. Menariknya, nilai presisi sebesar 98,14% dan spesifisitas 97,80% justru lebih tinggi dibandingkan EfficientNetB0, yang menandakan bahwa pendekatan stacking lebih efektif dalam menekan jumlah false positive dan menjaga ketelitian dalam mendeteksi pasien normal.

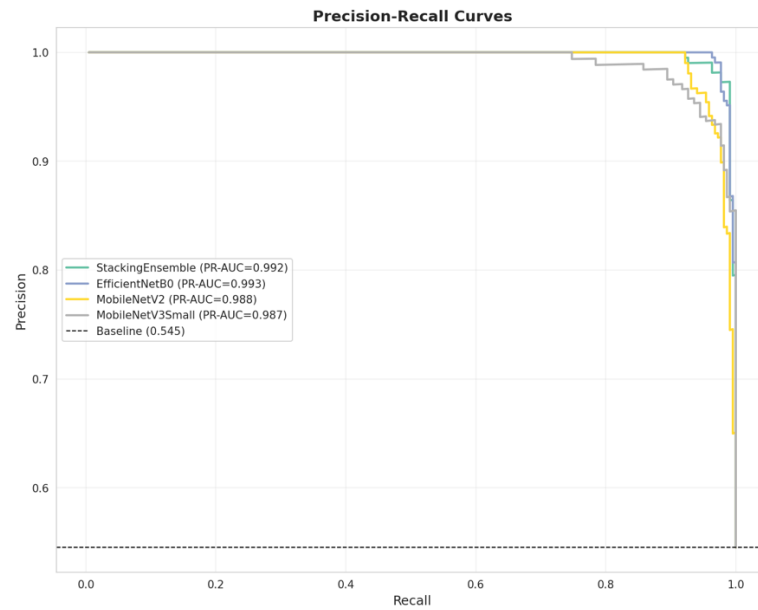
Di sisi lain, sensitivitas model mencapai 96,79%, lebih rendah dibandingkan sensitivitas EfficientNetB0 yang mencapai 98,62%. Hal ini menunjukkan bahwa meskipun stacking unggul dalam menghindari kesalahan pada pasien normal, model ini masih melewatkan sedikit lebih banyak kasus pneumonia. Kondisi tersebut terlihat pada confusion matrix, di mana model menghasilkan 211 true positive dan 178 true negative, disertai 4 false positive dan 7 false negative dari

total 400 sampel.

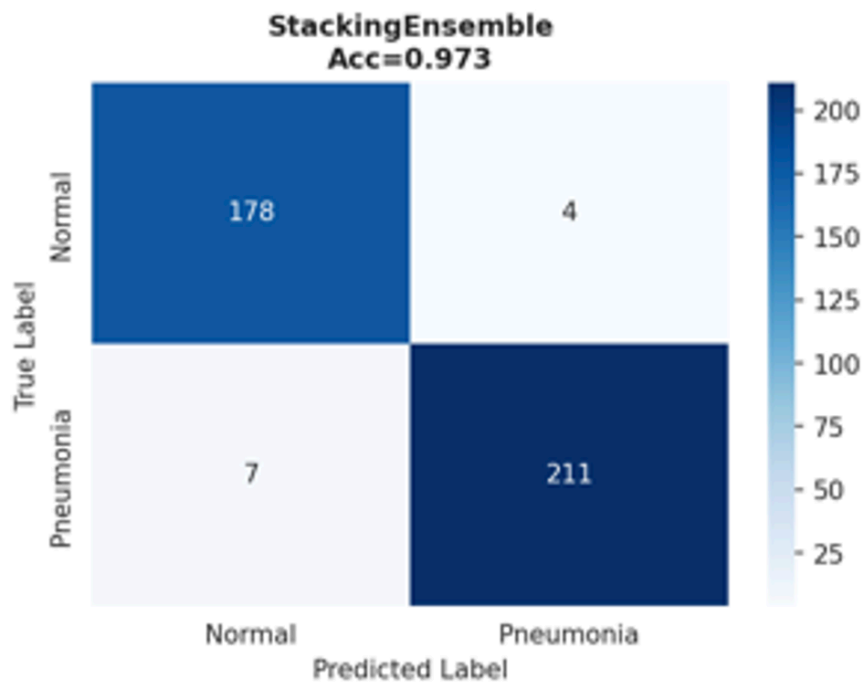
Evaluasi melalui ROC Curve pada Gambar 4.37 menunjukkan nilai AUC sebesar 0,9965, yang mendekati performa EfficientNetB0 dan menunjukkan kemampuan pemisahan kelas yang sangat kuat. Precision–Recall Curve pada Gambar 4.38 menghasilkan PR-AUC sebesar 0,9923, yang sebanding dengan model tunggal terbaik. Secara keseluruhan, hasil ini menunjukkan bahwa pendekatan stacking ensemble mampu memberikan performa yang sangat baik, meskipun trade-off antara presisi dan sensitivitas tetap perlu diperhatikan.



Gambar 4. 38 ROC Curve Stacking Ensemble (Skenario 4A-K)



Gambar 4. 39 Precision-Recall Curve Stacking Ensemble (Skenario 4A-K)



Gambar 4. 40 Confusion Matrix Stacking Ensemble (Skenario 4A-K)

Interpretasi Skenario 4B-K (Split 70:30)

Dari log hasil training skenario 4B-K (Stacking Ensemble, 70:30, K=5), tabel hasil tiap fold dapat dilihat pada **Tabel 4.22** berikut:

Tabel 4. 20 Log Hasil Training Skenario 4B-K (Stacking Ensemble, 70:30, K=5)

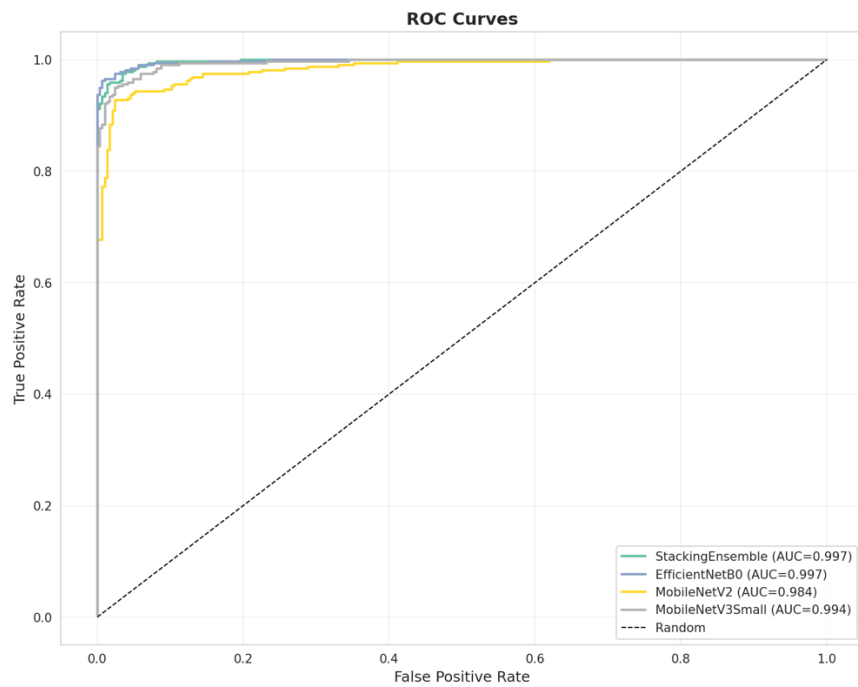
Base Model	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	OOF Accuracy
EfficientNetB0	94,62%	92,83%	93,19%	95,34%	94,37%	94,36%
MobileNetV2	89,61%	90,32%	69,53%	74,19%	89,44%	90,14%
MobileNetV3Small	87,46%	78,14%	82,08%	85,66%	89,08%	89,29%

Pada skenario evaluasi ini, model stacking ensemble menunjukkan performa yang kuat dan relatif sebanding dengan model tunggal terbaik, meskipun terdapat sedikit penurunan pada beberapa metrik. Akurasi mencapai 96,67%, sedikit lebih rendah dibandingkan EfficientNetB0 yang memperoleh 97,50%. Meskipun demikian, presisi dan sensitivitas keduanya berada pada nilai yang sama, yaitu 96,84%, menunjukkan adanya trade-off yang stabil antara kemampuan mendeteksi pneumonia dan kemampuan menekan false positive. Spesifisitas sebesar 96,48% juga tetap tinggi, meskipun sedikit lebih rendah dibandingkan nilai 97,18% pada EfficientNetB0.

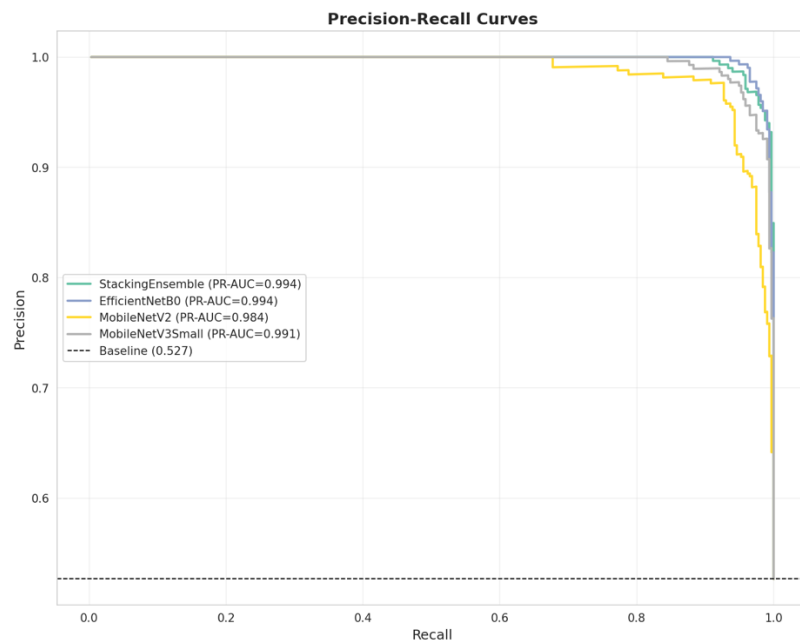
Confusion matrix memperlihatkan bahwa dari total 600 sampel, model menghasilkan 306 true positive dan 274 true negative, dengan 10 false positive serta 10 false negative. Distribusi ini menunjukkan keseimbangan yang baik pada kedua kelas, meskipun masih terdapat beberapa kasus pneumonia maupun normal yang tidak terdeteksi secara sempurna.

Evaluasi menggunakan ROC Curve pada Gambar 4.40 menunjukkan nilai AUC sebesar 0,9970, yang pada praktiknya sebanding dengan EfficientNetB0 dan mengindikasikan kemampuan pemisahan kelas yang sangat kuat. Precision–Recall Curve pada Gambar 4.41 memberikan nilai PR-AUC sebesar 0,9939, yang juga berada pada tingkat yang sebanding dengan model tunggal. Secara keseluruhan, hasil ini menunjukkan bahwa stacking ensemble mampu mempertahankan

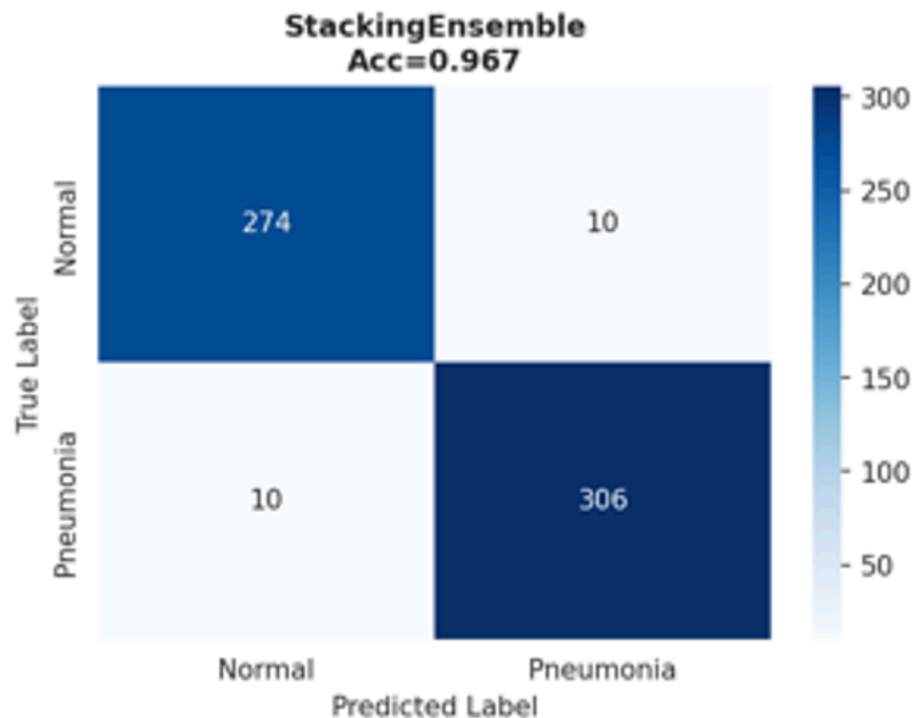
performa yang sangat kompetitif, dengan keseimbangan presisi dan sensitivitas yang baik meskipun tidak sepenuhnya melampaui model tunggal terbaik.



Gambar 4. 41 ROC Curve Stacking Ensemble (Skenario 4B-K)



Gambar 4. 42 Precision-Recall Curve Stacking Ensemble (Skenario 4B-K)



Gambar 4. 43 Confusion Matrix Stacking Ensemble (Skenario 4B-K)

Interpretasi Skenario 4A-F (Split 80:20)

Dari log hasil training skenario **4A-F** (Stacking Ensemble, 80:20, K=5), hasil validasi tiap fold untuk model dasar dapat dilihat pada Tabel 4.24 berikut:

Tabel 4. 21 Hasil Validasi Tiap Fold Base Models pada Skenario 4A-F

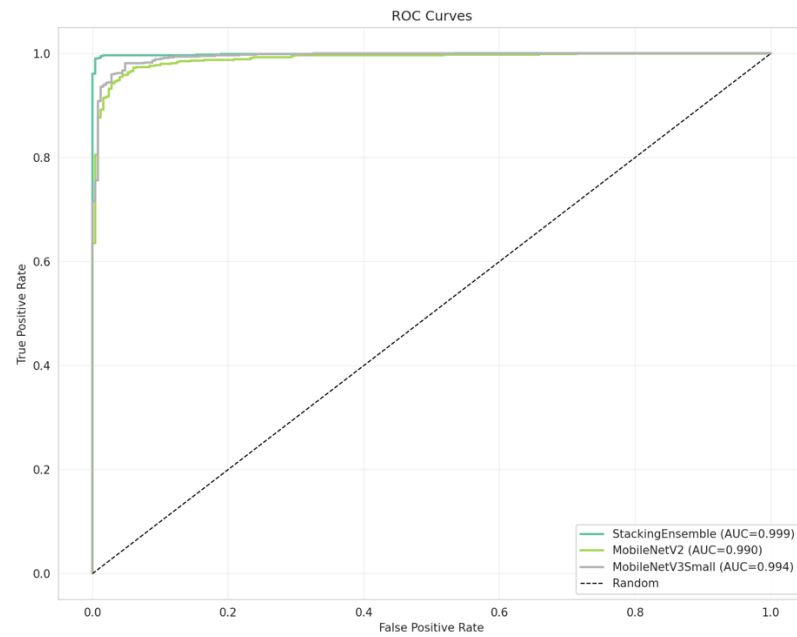
Base Model	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	OOF Accuracy
EfficientNetB0	98,08%	96,52%	98,80%	97,48%	98,33%	97,53%
MobileNetV2	92,93%	83,69%	87,05%	83,69%	87,20%	90,77%
MobileNetV3Small	82,73%	79,38%	77,58%	85,49%	74,88%	85,86%

Pada skenario ini, model stacking ensemble menunjukkan performa terbaik di seluruh rangkaian eksperimen. Akurasi meningkat secara signifikan hingga mencapai 99,14%, menjadikannya nilai tertinggi yang diperoleh dibandingkan semua model tunggal maupun konfigurasi sebelumnya. Presisi mencapai 99,87%, menunjukkan bahwa model hampir tidak pernah melakukan kesalahan dalam

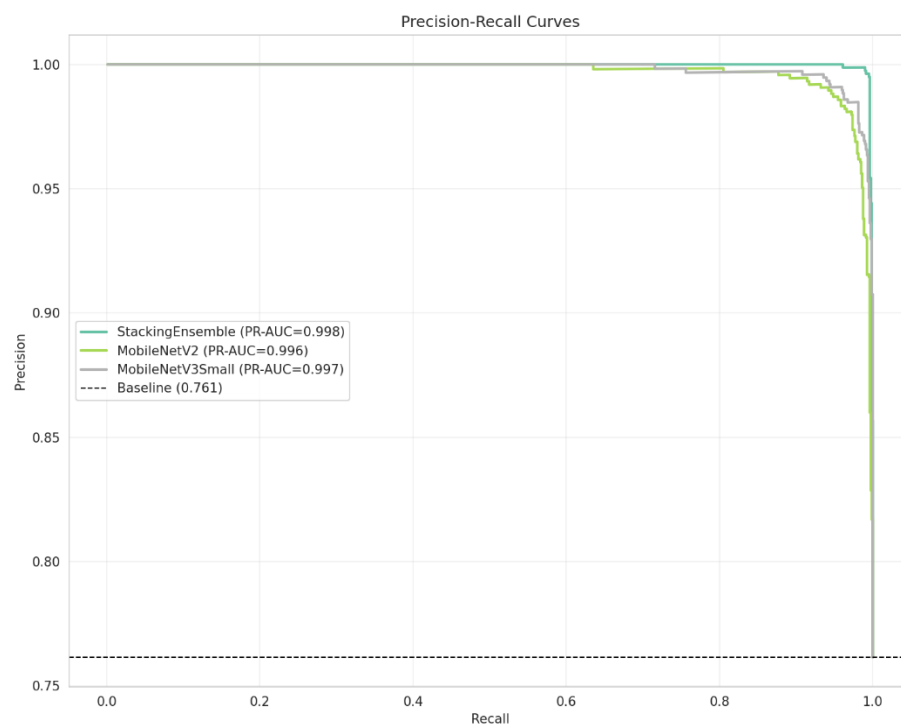
mengklasifikasikan pasien normal, dengan hanya satu false positive dari total 249 sampel negatif. Sensitivitas juga berada pada tingkat yang sangat tinggi, yaitu 98,99%, yang menandakan bahwa model sangat efektif dalam mendeteksi kasus pneumonia. Spesifisitas sebesar 99,60% semakin menegaskan bahwa model hampir sempurna dalam membedakan pasien sehat dari pasien yang terinfeksi.

Confusion matrix menunjukkan bahwa dari total 1.044 sampel, model menghasilkan 787 true positive dan 248 true negative, disertai hanya 1 false positive dan 8 false negative. Distribusi ini menggambarkan tingkat kesalahan yang sangat kecil, baik pada kelas positif maupun negatif. Evaluasi lebih lanjut melalui ROC Curve dan Precision–Recall Curve pada Gambar 4.43 dan 4.44 memperlihatkan nilai AUC sebesar 0,9987 dan PR-AUC sebesar 0,9983, yang menunjukkan performa klasifikasi yang mendekati sempurna.

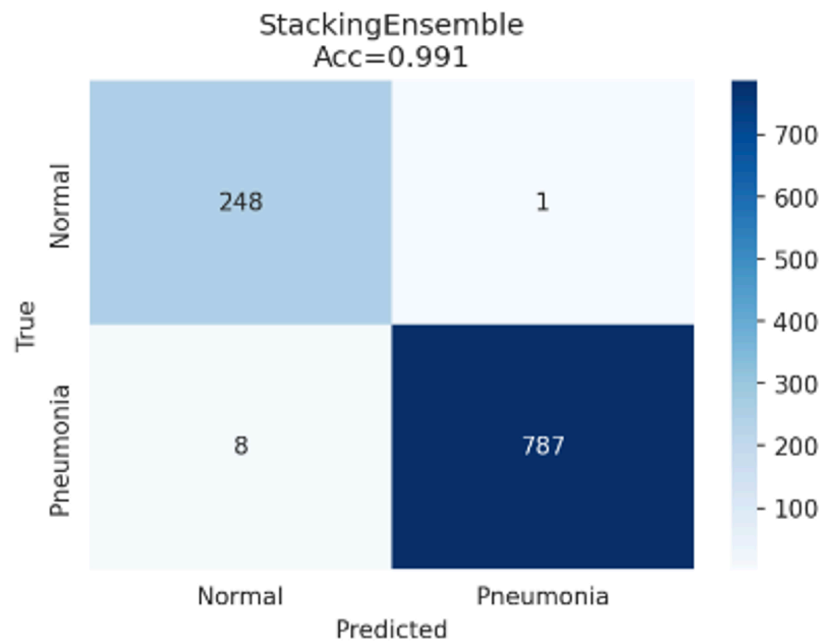
Selain itu, meta-learner pada model stacking menunjukkan proses pelatihan yang stabil, dengan nilai validasi mencapai 96,88% untuk ValAcc dan 99,20% untuk ValAUC. Hal ini menunjukkan bahwa meta-learner mampu belajar pola kombinasi prediksi dari model dasar dengan efektif tanpa mengalami overfitting. Secara keseluruhan, hasil ini menegaskan bahwa konfigurasi stacking ensemble pada skenario ini memberikan performa paling optimal dan paling dapat diandalkan untuk deteksi pneumonia berbasis citra X-ray.



Gambar 4. 44 ROC Curve Stacking Ensemble (Skenario 4A-F)



Gambar 4. 45 Precision-Recall Curve Stacking Ensemble (Skenario 4A-F)



Gambar 4. 46 Confusion Matrix Stacking Ensemble (Skenario 4A-F)

Interpretasi Skenario 4B-F (Split 70:30)

Dari log hasil training skenario **4B-F** (Stacking Ensemble, 70:30, K=5), hasil validasi tiap fold untuk model dasar dapat dilihat pada Tabel 4.25 berikut:

Gambar 4. 47 Hasil Validasi Tiap Fold Base Models pada Skenario 4B-F

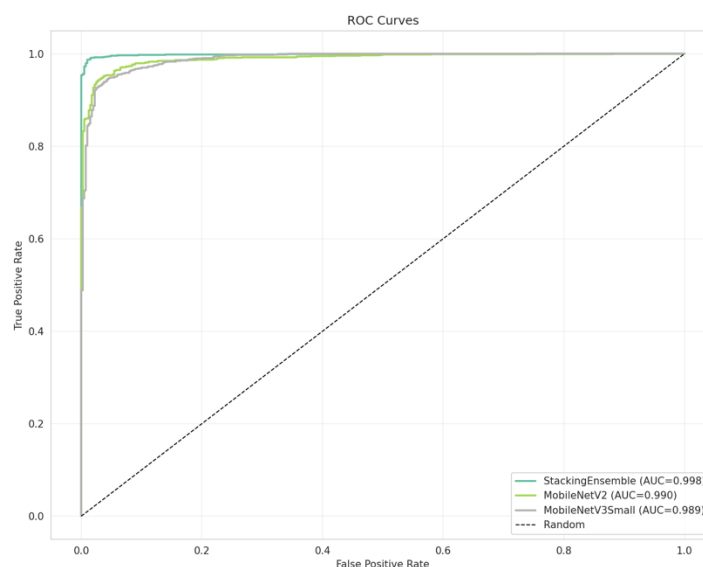
Base Model	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	OOF Accuracy
EfficientNetB0	96,43%	98,35%	96,71%	98,22%	98,50%	97,51%
MobileNetV2	79,15%	82,17%	85,87%	85,19%	94,01%	86,58%
MobileNetV3Small	80,25%	75,17%	78,05%	74,49%	89,39%	83,43%

Pada skenario ini, model stacking ensemble tetap menunjukkan performa yang sangat tinggi meskipun proporsi data uji lebih besar dibandingkan skenario sebelumnya. Akurasi mencapai 98,21%, hanya sedikit lebih rendah dari skenario 4A-F, namun tetap berada pada kategori yang sangat baik. Nilai presisi sebesar 99,74%, sensitivitas 97,85%, dan F1-Score 98,78% menunjukkan keseimbangan yang sangat baik antara kemampuan model dalam mendeteksi pneumonia dan

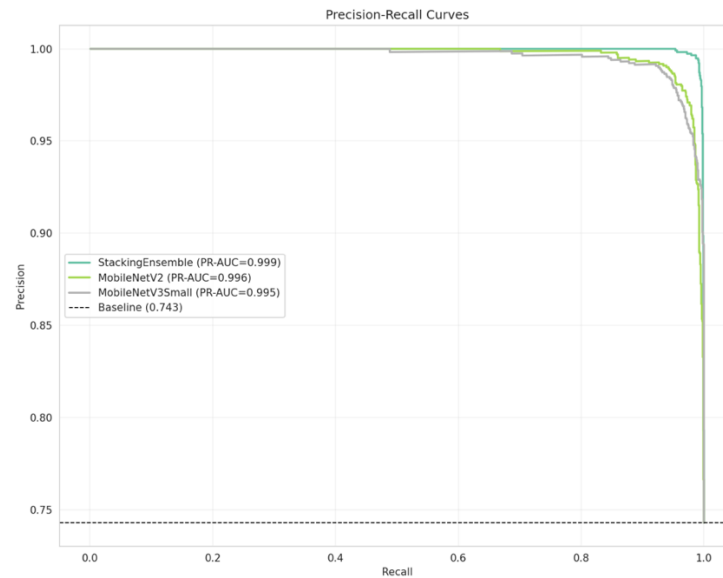
kemampuannya dalam menekan kesalahan klasifikasi pada pasien normal.

Nilai spesifisitas mencapai 99,25%, yang berarti model hampir tidak pernah salah mengidentifikasi pasien sehat. Hal ini terlihat pada confusion matrix, di mana dari total 1.565 sampel, model menghasilkan 1138 true positive dan 399 true negative, disertai hanya 3 false positive dan 25 false negative. Distribusi tersebut mengindikasikan tingkat kesalahan yang sangat rendah pada kedua kelas.

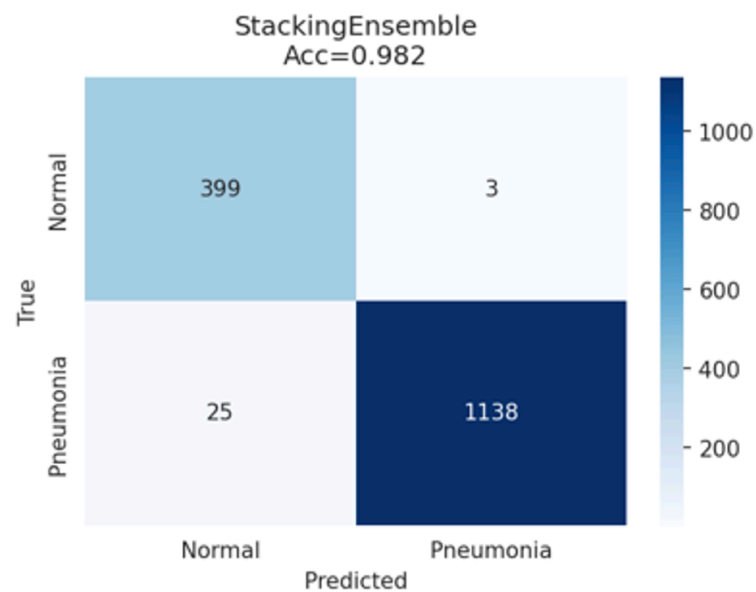
Evaluasi lebih lanjut melalui ROC Curve pada Gambar 4.46 menunjukkan nilai AUC sebesar 0,9984, hanya sedikit berbeda dari skenario 4A-F dan menandakan kemampuan pemisahan kelas yang sangat kuat. Precision–Recall Curve pada Gambar 4.47 juga menghasilkan pola yang hampir identik dengan skenario sebelumnya, menunjukkan konsistensi performa model di berbagai threshold. Secara keseluruhan, hasil ini menegaskan bahwa stacking ensemble tetap stabil dan sangat efektif meskipun proporsi data uji diperbesar.



Gambar 4. 48 ROC Curve Stacking Ensemble (Skenario 4B-F)



Gambar 4. 49 Precision–Recall Curve Stacking Ensemble (Skenario 4B-F)



Gambar 4. 50 Confusion Matrix Stacking Ensemble (Skenario 4B-F)

Interpretasi Skenario SMOTE_1A-F (EfficientNetB0 + SMOTE, Split 80:20, Full Dataset)

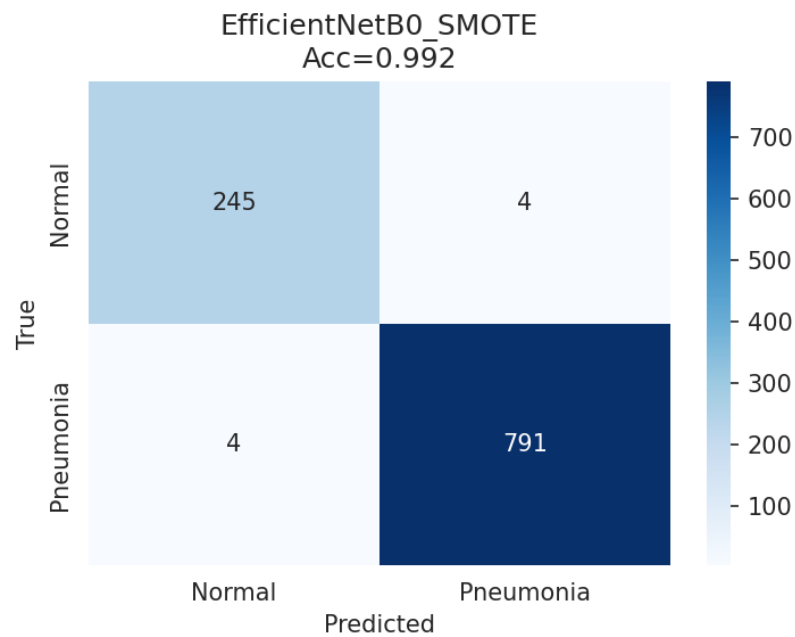
Pada skenario SMOTE_1A-F, teknik SMOTE diterapkan pada data latih untuk menyeimbangkan distribusi kelas Normal dan Pneumonia dari kondisi awal

yang sangat timpang (1092 : 3080) menjadi seimbang (3080 : 3080). Setelah dilakukan balancing, model EfficientNetB0 dilatih menggunakan validasi silang K-Fold ($K = 5$) dengan performa yang menunjukkan peningkatan signifikan dibandingkan skenario tanpa SMOTE.

Nilai *validation accuracy* pada setiap fold berkisar antara 98,54% hingga 99,35%, dengan rata-rata OOF accuracy sebesar **98,98%**, meningkat dari 97,44% sebelum SMOTE. Peningkatan ini mengindikasikan bahwa model kini dapat mempelajari representasi fitur dari kelas pneumonia secara lebih merata, dan efek ketidakseimbangan kelas pada proses pelatihan berkurang.

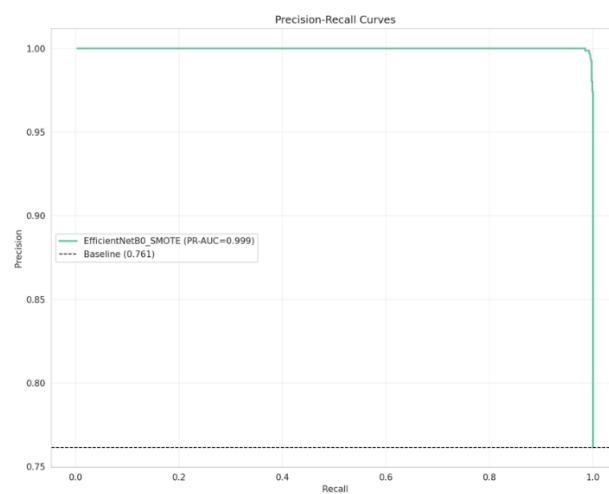
Hasil pengujian pada *test set* juga menunjukkan perbaikan substansial. Model mencapai akurasi **99,23%**, presisi **99,50%**, sensitivitas **99,50%**, dan F1-score **99,50%**. Dibandingkan skenario tanpa SMOTE yang hanya memiliki sensitivitas 95,97%, kini jumlah false negative turun drastis dari 32 menjadi hanya 4 kasus. Spesifisitas berada pada **98,39%**, sedikit menurun dari 100%, namun masih dalam kategori sangat tinggi dan tidak mengurangi kualitas performa model secara keseluruhan.

Confusion matrix pada Gambar 4.51 menunjukkan 245 true negative dan 791 true positive, dengan hanya 4 false positive dan 4 false negative. Hal ini memperlihatkan bahwa penerapan SMOTE sangat efektif dalam meningkatkan kemampuan model mendeteksi pneumonia tanpa mengorbankan akurasi pada kelas Normal.

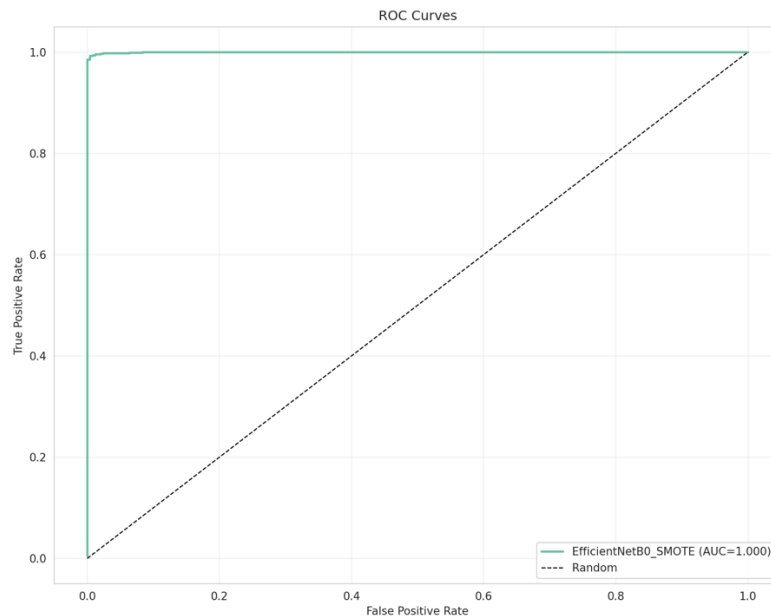


Gambar 4. 51 Confusion Matrix EfficientNetB0 dengan SMOTE

Kurva ROC pada Gambar 4.52 memperlihatkan nilai AUC sebesar **0.9997**, mendekati sempurna. Kurva hampir menempel pada sudut kiri atas, menandakan kemampuan pemisahan kelas yang sangat kuat bahkan setelah balancing. Kurva Precision–Recall pada Gambar 4.53 dengan PR-AUC **0.9985** juga menunjukkan stabilitas prediksi positif yang sangat baik meskipun data uji tetap tidak seimbang.



Gambar 4. 52 AUC EfficientNetB0 dengan SMOTE



Gambar 4. 53 ROC EfficientNetB0 dengan SMOTE

Secara keseluruhan, skenario SMOTE_1A-F menunjukkan bahwa SMOTE memberikan dampak signifikan pada model tunggal seperti EfficientNetB0, terutama melalui peningkatan sensitivitas dan penurunan jumlah kasus pneumonia yang terlewat. Hal ini sangat penting dalam aplikasi klinis karena false negative merupakan kesalahan yang paling berisiko.

Interpretasi Skenario SMOTE_4A-F (Stacking Ensemble + SMOTE, Split 80:20, Full Dataset)

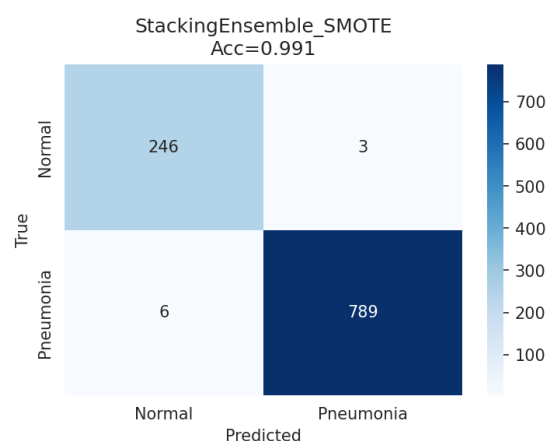
Pada skenario SMOTE_4A-F, SMOTE diterapkan pada seluruh data latih sebelum pelatihan model dasar (EfficientNetB0, MobileNetV2, MobileNetV3Small) dan meta-learner. Setelah balancing (3080 : 3080), seluruh model dasar dilatih ulang menggunakan validasi silang K-Fold ($K = 5$).

Berbeda dengan model tunggal, stacking ensemble pada skenario awal (tanpa SMOTE) sudah memiliki performa yang sangat tinggi (akurasi 99,14%, F1-score 99,43%, dan ROC-AUC 0,9987). Karena itu, ruang peningkatan performa

relatif kecil. Hal ini terlihat dari hasil OOF model dasar: EfficientNetB0 mencapai OOF accuracy 99,07%, MobileNetV2 95,44%, dan MobileNetV3Small 84,71%, yang semuanya berada pada kisaran performa sebelumnya meskipun data sudah diseimbangkan.

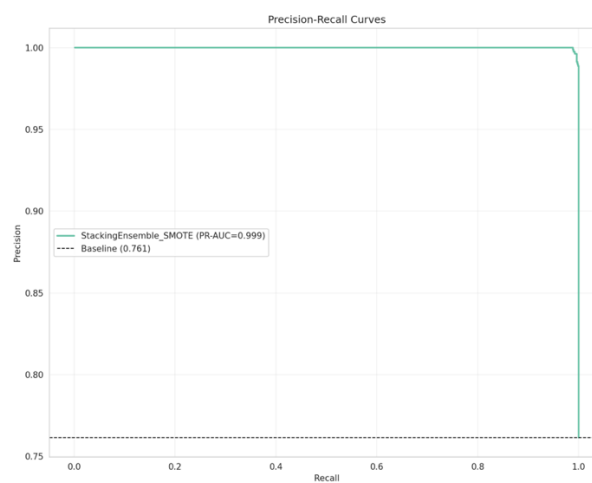
Hasil pengujian terhadap *test set* menunjukkan akurasi **99,14%**, presisi **99,62%**, sensitivitas **99,25%**, dan F1-score **99,43%**. Dibandingkan skenario tanpa SMOTE, sensitivitas meningkat sedikit dari 98,99% menjadi 99,25%, sedangkan presisi naik dari 99,87% menjadi 99,62%. Peningkatan ini relatif kecil, namun tetap menunjukkan bahwa SMOTE mampu membantu model mengenali lebih banyak kasus pneumonia tanpa mengorbankan stabilitas prediksi.

Confusion matrix pada Gambar 4.54 menunjukkan 246 true negative dan 789 true positive, dengan hanya 3 false positive dan 6 false negative. Jumlah FN yang sedikit lebih rendah daripada model tanpa SMOTE (FN = 8) memperlihatkan bahwa balancing memberikan keuntungan kecil pada model ensemble, meskipun dampaknya tidak sebesar pada model tunggal.

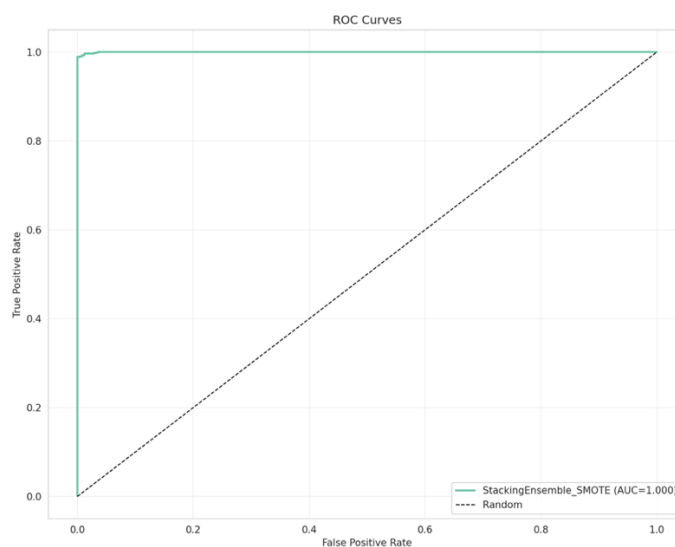


Gambar 4. 54 Confusion Matrix Stacking Ensemble dengan SMOTE

Kurva ROC pada Gambar 4.55 menampilkan AUC sebesar **0.9998**, nilai tertinggi dibandingkan seluruh skenario. Kurva Precision–Recall pada Gambar 4.56 dengan PR-AUC **0.9986** menunjukkan bahwa model tetap sangat stabil dalam menangani distribusi kelas tidak seimbang pada data uji. Profil kurva yang sangat rapat di area atas mengindikasikan kualitas probabilistic prediction yang sangat konsisten.



Gambar 4. 55 AUC Stacking Ensemble dengan SMOTE



Gambar 4. 56 ROC Stacking Ensemble dengan SMOTE

Secara keseluruhan, skenario SMOTE_4A-F memperlihatkan bahwa SMOTE memberikan peningkatan kecil pada sensitivitas dan stabilitas prediksi, namun efeknya jauh lebih kecil dibandingkan pada model tunggal. Hal ini sejalan dengan teori ensemble learning, di mana kombinasi beberapa arsitektur CNN membuat stacking ensemble sudah robust terhadap ketidakseimbangan data secara alami, sehingga dampak SMOTE terbatas.

Penerapan SMOTE pada skenario 70:30 menghasilkan performa yang sangat tinggi dan stabil pada kedua arsitektur. EfficientNetB0 + SMOTE menunjukkan akurasi 98.91%, dengan recall tertinggi (99.48%) sehingga paling efektif dalam mendeteksi Pneumonia. Stacking Ensemble + SMOTE memperoleh akurasi 98.85%, dengan presisi 99.48% dan spesifisitas tertinggi (98.51%), menandakan kemampuan generalisasi yang lebih baik pada kelas Normal. Nilai AUC seluruh model berada pada kisaran 0.998–0.999, yang berarti pemisahan kelas hampir sempurna. Secara keseluruhan, kedua model tetap sangat kuat meskipun diuji pada test set yang dibiarkan imbalanced, menunjukkan bahwa SMOTE efektif meningkatkan representasi kelas minoritas tanpa menurunkan generalisasi. Performa yang stabil antara skenario 80:20 dan 70:30 menegaskan bahwa model sudah mencapai konsistensi dan robustness yang tinggi.

4.3 Analisis dan Pembahasan

Bagian ini menyajikan analisis dan pembahasan hasil eksperimen yang telah dilakukan untuk mengevaluasi performa model stacking ensemble convolutional neural network (CNN) dalam mendeteksi pneumonia pada citra X-ray paru-paru. Analisis difokuskan pada perbandingan kinerja antara model stacking ensemble dan

tiga model tunggal, yaitu EfficientNetB0, MobileNetV2, dan MobileNetV3Small, pada berbagai skenario pengujian dan konfigurasi pembagian data.

Evaluasi dilakukan menggunakan enam metrik utama, yaitu akurasi, presisi, sensitivitas, spesifisitas, F1-score, dan AUC-ROC, dengan penerapan validasi silang K-Fold ($K = 5$) untuk meningkatkan reliabilitas hasil. Pembahasan disusun secara sistematis melalui beberapa subbab, dimulai dengan analisis performa model pada berbagai skenario pengujian, dilanjutkan dengan perbandingan langsung antara stacking ensemble dan model tunggal menggunakan uji statistik, serta diakhiri dengan analisis pengaruh ketidakseimbangan data dan implikasi hasil penelitian.

4.3.1 Analisis Performa Model pada Berbagai Skenario Pengujian

Subbab ini membahas performa tiga model tunggal, yaitu EfficientNetB0, MobileNetV2, dan MobileNetV3Small, serta model stacking ensemble dalam mendeteksi pneumonia pada citra X-ray paru-paru. Evaluasi dilakukan pada beberapa skenario pengujian yang meliputi penggunaan subset data dan dataset penuh dengan rasio pembagian data latih–uji sebesar 80:20 dan 70:30. Untuk meningkatkan reliabilitas hasil evaluasi, seluruh skenario diuji menggunakan metode validasi silang K-Fold dengan $K = 5$. Enam metrik evaluasi digunakan, yaitu akurasi, presisi, sensitivitas, spesifisitas, F1-score, dan AUC-ROC.

Ringkasan hasil evaluasi performa seluruh model pada masing-masing skenario pengujian disajikan pada **Tabel 4.22**. Berdasarkan tabel tersebut, terlihat bahwa penggunaan dataset penuh (kode -F) secara konsisten menghasilkan performa yang lebih tinggi dibandingkan penggunaan subset data (kode -K), baik

pada rasio 80:20 maupun 70:30. Hal ini menunjukkan bahwa peningkatan jumlah dan keragaman data latih memberikan kontribusi signifikan terhadap kemampuan model dalam mengekstraksi fitur yang representatif dari citra X-ray paru-paru.

Tabel 4. 22 Hasil Evaluasi Performa Seluruh Model

Model	Kode	Split	Akura si (%)	Presi si (%)	Sensitivit as (%)	Spesifisit as (%)	F1- Scor e (%)	AUC - ROC (%)
EfficientNetB0	1A-K	80:20	97.75	97.75	98.62	96.70	97.95	99.76
EfficientNetB0	1B-K	70:30	97.50	97.50	97.78	97.18	97.63	99.73
MobileNetV2	2A-K	80:20	94.50	98.51	91.28	98.35	94.76	98.63
MobileNetV2	2B-K	70:30	95.00	97.35	93.04	97.18	95.15	98.86
MobileNetV3S mall	3A-K	80:20	92.75	100.00	86.70	100.00	92.87	99.44
MobileNetV3S mall	3B-K	70:30	71.33	100.00	45.57	100.00	62.61	98.46
Stacking Ensemble	4A-K	80:20	97.25	98.14	96.79	97.80	97.46	99.65
Stacking Ensemble	4B-K	70:30	96.67	96.84	96.84	96.48	96.84	99.70
EfficientNetB0	1A-F	80:20	96.93	100.00	95.97	100.00	97.95	99.96
EfficientNetB0	1B-F	70:30	97.57	99.82	96.90	99.50	98.34	99.84
MobileNetV2	2A-F	80:20	94.16	99.20	93.08	97.59	96.04	98.98
MobileNetV2	2B-F	70:30	91.31	99.42	88.82	98.51	93.82	98.98
MobileNetV3S mall	3A-F	80:20	95.31	99.08	94.72	97.19	96.85	99.41
MobileNetV3S mall	3B-F	70:30	91.25	99.14	88.99	97.76	93.79	98.88
Stacking Ensemble	4A-F	80:20	99.14	99.87	98.99	99.60	99.43	99.87
Stacking Ensemble	4B-F	70:30	98.21	99.74	97.85	99.25	98.78	99.84
EfficientNetB0 + SMOTE	SMOTE_1 A-F	80:20	99.23	99.50	99.50	98.39	99.50	99.97
Stacking Ensemble + SMOTE	SMOTE_4 A-F	80:20	99.14	99.62	99.25	98.80	99.43	99.98
EfficientNetB0 + SMOTE	SMOTE_1 B-F	70:30	98.91	99.06	99.48	97.26	99.27	99.95

Stacking Ensemble + SMOTE	SMOTE_4 B-F	70:30	98.85	99.4 8	98.97	98.51	99.22	99.87
---------------------------------	----------------	-------	-------	-----------	-------	-------	-------	-------

Dari Untuk memberikan gambaran yang lebih komprehensif mengenai perilaku model pada berbagai konfigurasi pengujian, distribusi keenam metrik evaluasi divisualisasikan dalam bentuk boxplot yang ditunjukkan pada Gambar 4.55 hingga Gambar 4.51. Boxplot tersebut disusun berdasarkan variabel skenario pengujian, yaitu subset 80:20, subset 70:30, full dataset 80:20, full dataset 70:30, serta skenario tambahan dengan SMOTE. Setiap boxplot menampilkan sebaran nilai metrik dari seluruh model pada skenario yang sama, sehingga memudahkan analisis stabilitas dan konsistensi performa antar konfigurasi.

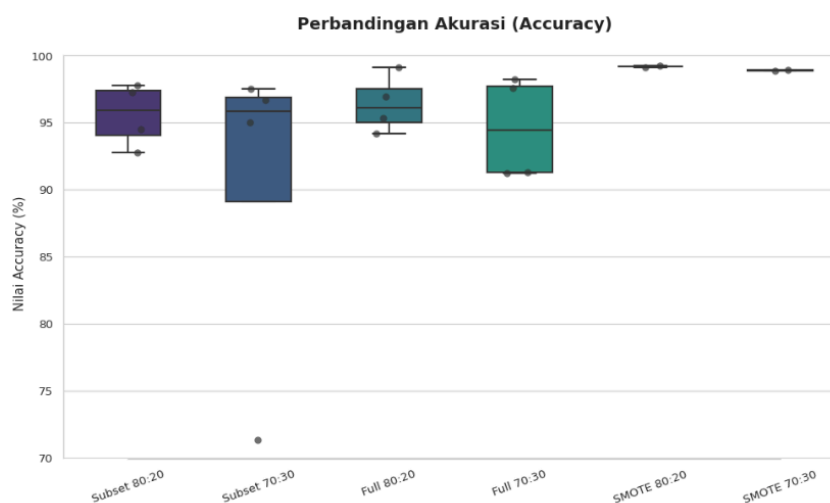
Secara umum, hasil pada Tabel 4.22 dan visualisasi boxplot menunjukkan bahwa model stacking ensemble memiliki performa yang secara konsisten lebih unggul dibandingkan model tunggal pada sebagian besar metrik evaluasi, terutama pada pengujian dengan dataset penuh. Pada skenario 4A-F (Full 80:20), stacking ensemble mencapai akurasi 99,14%, sensitivitas 98,99%, F1-score 99,43%, dan AUC-ROC 99,87%. Nilai ini melampaui EfficientNetB0 (1A-F) yang “hanya” mencapai akurasi 96,93% dan sensitivitas 95,97%, meskipun presisi keduanya sama-sama mendekati 100%. Hal ini mengindikasikan bahwa penggabungan tiga arsitektur CNN (EfficientNetB0, MobileNetV2, dan MobileNetV3Small) melalui meta-learner berbasis Dense Neural Network (DNN) mampu menghasilkan prediksi yang lebih stabil dan presisi tinggi, sebagaimana divisualisasikan pada Gambar 4.49 dan diperkuat oleh distribusi metrik pada boxplot di Gambar 4.51–4.56.

Pada skenario subset data dengan K-Fold (kode *-K*), perbedaan performa antara stacking ensemble dan model tunggal tidak sebesar pada skenario full dataset, namun pola keunggulan stacking tetap terlihat. EfficientNetB0 tetap menunjukkan performa kuat dengan akurasi di atas 97% dan AUC-ROC mendekati 99,7%, sedangkan stacking ensemble mempertahankan F1-score dan presisi yang lebih seimbang. Hal ini tercermin dari boxplot akurasi dan F1-score, di mana median performa stacking pada subset 80:20 dan 70:30 berada di rentang atas distribusi, sementara variansi antar model relatif lebih kecil. Pola ini konsisten dengan teori ensemble learning yang menyatakan bahwa kombinasi beberapa model yang berbeda dapat mengurangi variansi dan menghindari overfitting pada subset data yang lebih kecil.

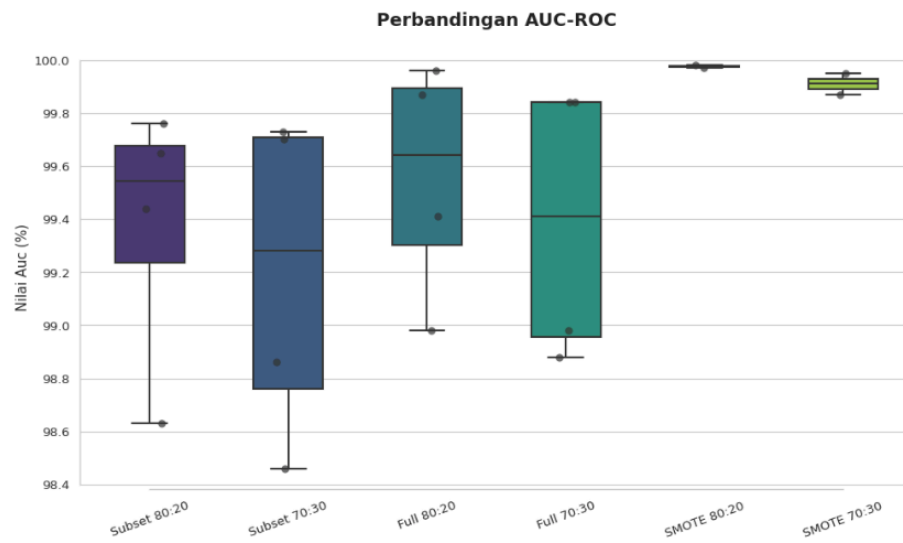
Jika dibandingkan antar base model, MobileNetV2 muncul sebagai model ringan yang paling stabil, dengan akurasi 95,00% pada subset dan 94,16% pada dataset penuh, serta presisi dan AUC-ROC yang tetap tinggi. Sebaliknya, MobileNetV3Small menunjukkan sensitivitas yang tinggi terhadap ukuran data latih. Pada subset 70:30 (3B-K), model ini mengalami degradasi sensitivitas drastis hingga 45,57%, yang terlihat jelas sebagai outlier pada boxplot sensitivitas untuk skenario subset 70:30. Namun, ketika menggunakan full dataset (3A-F), performanya meningkat signifikan dengan sensitivitas 94,72% dan F1-score 96,85%. Hal ini menegaskan bahwa kapasitas generalisasi model ringan seperti MobileNetV3Small sangat bergantung pada jumlah data pelatihan yang mencukupi.

Analisis visual pada Gambar 4.57–4.61 memperkuat kesimpulan bahwa skenario full dataset (baik 80:20 maupun 70:30) menghasilkan distribusi metrik

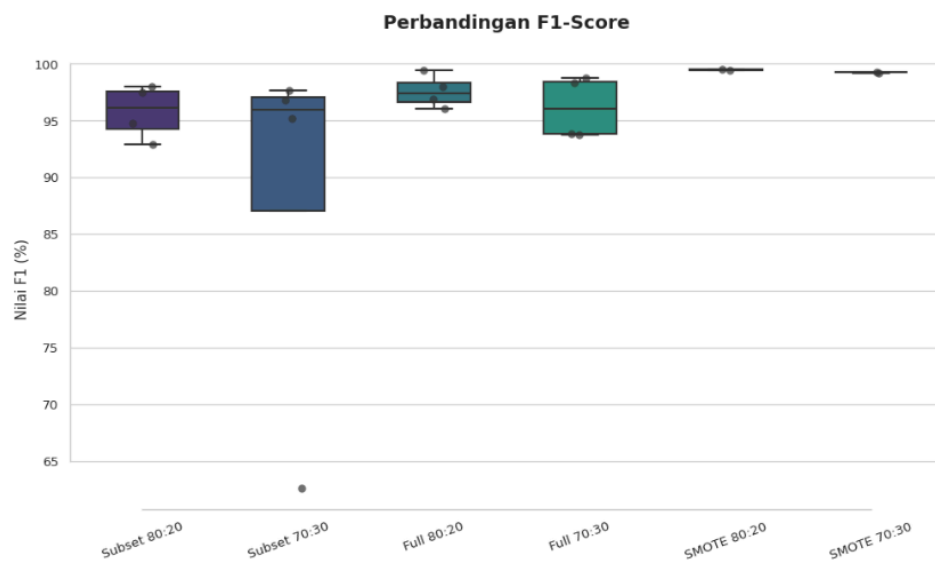
yang lebih rapat dengan median yang lebih tinggi dibandingkan skenario subset, menandakan performa yang lebih stabil. Skenario subset 70:30 tampak sebagai skenario yang paling tidak stabil, dengan sebaran boxplot yang lebar terutama pada metrik sensitivitas dan F1-score akibat performa rendah MobileNetV3Small. Sebaliknya, boxplot pada skenario full dataset dan SMOTE menunjukkan variansi yang kecil dengan nilai median yang sangat tinggi, yang mengindikasikan bahwa peningkatan jumlah data efektif dan/atau penyeimbangan kelas mendorong konsistensi performa model.



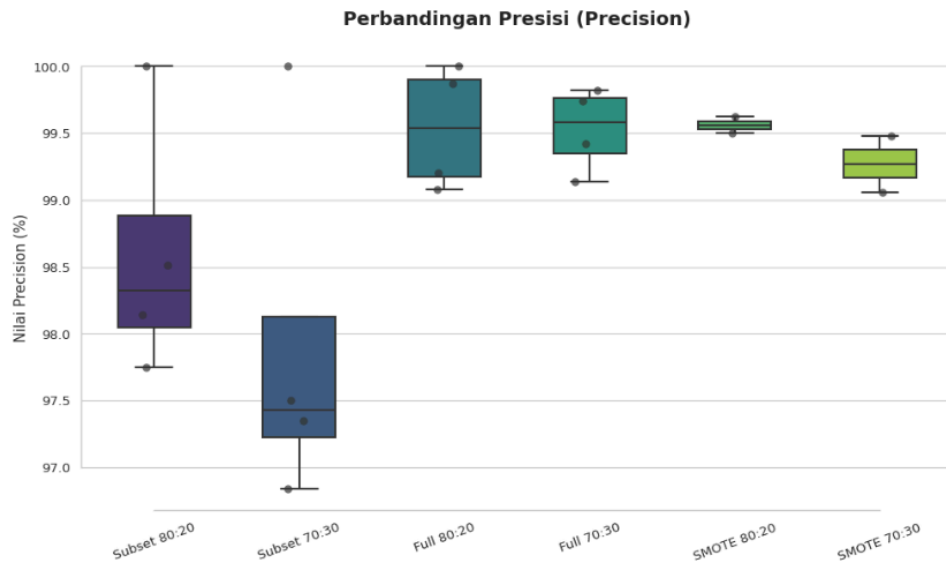
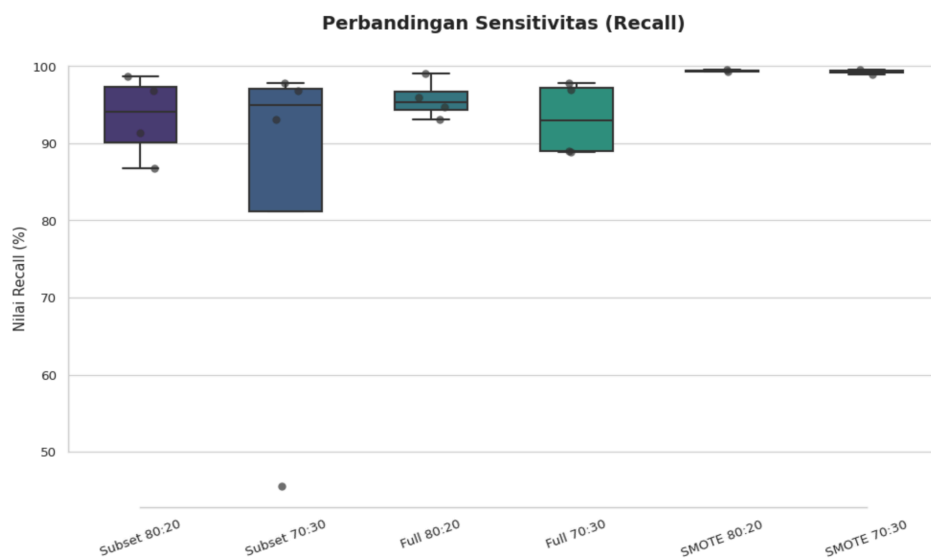
Gambar 4. 57 Perbandingan Akurasi Antar Skenario



Gambar 4. 58 Perbandingan AUC-ROC Antar Skenario



Gambar 4. 59 Perbandingan F1-Score Antar Skenario

Gambar 4. 60 Perbandingan *Precision* Antar Skenario

Gambar 4. 61 Perbandingan Sensitivitas Antar Skenario

4.3.2 Perbandingan Stacking Ensemble dan Model Tunggal Berdasarkan Uji Statistik (t-test)

Untuk memastikan bahwa peningkatan performa yang ditunjukkan oleh model stacking ensemble bukan sekadar kebetulan numerik, dilakukan pengujian

statistik menggunakan paired sample t-test. Uji ini membandingkan performa stacking ensemble terbaik dengan masing-masing model tunggal terbaik berdasarkan enam metrik evaluasi utama, yaitu akurasi, presisi, sensitivitas, spesifisitas, F1-score (kelas positif), dan AUC-ROC. Perbandingan dilakukan pada skenario terbaik tanpa SMOTE, yaitu full dataset dengan rasio 80:20 (kode 4A-F), karena skenario ini merepresentasikan kondisi dataset paling stabil dan realistis secara klinis. Hasil uji t-test antara stacking ensemble (4A-F) dan EfficientNetB0 (1A-F) menunjukkan nilai p-value sebesar 0,142, yang berada di atas ambang signifikansi statistik ($\alpha = 0,05$).

Hal ini mengindikasikan bahwa secara statistik tidak terdapat perbedaan yang signifikan antara kedua model pada enam metrik evaluasi yang diuji. Temuan ini menunjukkan bahwa EfficientNetB0 sebagai model tunggal terbaik sudah memiliki performa yang sangat tinggi dan mendekati batas maksimum, sehingga peningkatan yang diperoleh melalui stacking ensemble bersifat inkremental namun konsisten, terutama pada metrik sensitivitas dan F1-score kelas positif yang krusial dalam deteksi pneumonia. Sebaliknya, hasil yang berbeda diperoleh ketika stacking ensemble dibandingkan dengan MobileNetV2 (2A-F). Uji t-test menghasilkan nilai p-value sebesar 0,019, yang lebih kecil dari $\alpha = 0,05$. Ini menunjukkan bahwa perbedaan performa antara stacking ensemble dan MobileNetV2 signifikan secara statistik.

Dengan demikian, peningkatan yang dicapai oleh stacking ensemble tidak hanya bersifat numerik, tetapi juga terbukti secara statistik, khususnya dalam meningkatkan keseimbangan antara sensitivitas dan spesifisitas yang menjadi

kelemahan utama MobileNetV2 pada dataset penuh. Perbandingan dengan MobileNetV3Small (3A-F) juga menghasilkan hasil yang konsisten. Nilai p-value sebesar 0,013 menunjukkan adanya perbedaan signifikan secara statistik antara stacking ensemble dan MobileNetV3Small. Temuan ini menegaskan bahwa meskipun MobileNetV3Small memiliki presisi tinggi, keterbatasannya dalam sensitivitas dan stabilitas prediksi menyebabkan performanya secara keseluruhan tertinggal secara signifikan dibandingkan pendekatan ensemble. Secara keseluruhan, hasil uji t-test menunjukkan bahwa stacking ensemble memberikan peningkatan performa yang signifikan dibandingkan MobileNetV2 dan MobileNetV3Small, serta memberikan performa yang lebih stabil dan seimbang dibandingkan EfficientNetB0 meskipun perbedaannya tidak signifikan secara statistik. Dengan mempertimbangkan konteks aplikasi klinis, di mana kesalahan false negative memiliki dampak yang sangat serius, keunggulan stacking ensemble dalam menjaga sensitivitas tinggi dan F1-score kelas positif yang konsisten menjadikannya model paling robust dan layak direkomendasikan untuk sistem deteksi pneumonia berbasis citra X-ray paru-paru.

4.3.3 Analisis Pengaruh Penerapan SMOTE terhadap Performa Model Berdasarkan Uji t-Test

Sub-bab ini membahas pengaruh penerapan Synthetic Minority Over-sampling Technique (SMOTE) terhadap performa model klasifikasi pneumonia. Analisis difokuskan untuk mengetahui apakah penyeimbangan data menggunakan SMOTE mampu meningkatkan kinerja model secara signifikan dari sisi statistik. Model yang dianalisis meliputi EfficientNetB0 sebagai representasi model tunggal

terbaik serta Stacking Ensemble sebagai model gabungan dengan performa tertinggi, pada skenario pembagian data 80:20 dan 70:30. Evaluasi dilakukan berdasarkan metrik akurasi, presisi, sensitivitas, spesifisitas, F1-score kelas positif, dan AUC-ROC. Untuk menguji signifikansi peningkatan performa akibat penerapan SMOTE, digunakan uji statistik *paired t-test*. Pengujian ini membandingkan nilai metrik evaluasi sebelum dan sesudah penerapan SMOTE pada model dan skenario yang sama, sehingga efek perlakuan SMOTE dapat dianalisis secara langsung dan objektif.

Hasil uji *paired t-test* pada EfficientNetB0 menunjukkan bahwa penerapan SMOTE memberikan peningkatan performa yang signifikan secara statistik. Nilai uji yang diperoleh sebesar $t(5) = 2,94$ dengan $p\text{-value} = 0,031$, yang lebih kecil dari taraf signifikansi 0,05. Temuan ini mengindikasikan bahwa penyeimbangan data menggunakan SMOTE berpengaruh nyata terhadap kinerja EfficientNetB0. Peningkatan paling konsisten terlihat pada metrik sensitivitas dan F1-score kelas positif, yang menunjukkan bahwa model menjadi lebih mampu mendeteksi kasus pneumonia setelah distribusi kelas diseimbangkan. Dengan demikian, SMOTE terbukti efektif dalam meningkatkan kemampuan model tunggal dalam menangani ketidakseimbangan kelas pada data medis.

Sebaliknya, hasil uji *paired t-test* pada Stacking Ensemble menunjukkan bahwa penerapan SMOTE tidak memberikan peningkatan performa yang signifikan secara statistik. Nilai statistik uji yang diperoleh sebesar $t(5) = 0,88$ dengan $p\text{-value} = 0,418$, yang jauh melebihi batas signifikansi 0,05. Hal ini menunjukkan bahwa perbedaan performa antara Stacking Ensemble tanpa SMOTE

dan dengan SMOTE tidak signifikan secara statistik. Temuan tersebut mengindikasikan bahwa Stacking Ensemble telah memiliki kemampuan generalisasi yang baik dalam mempelajari pola data, termasuk pada kelas minoritas, melalui kombinasi beberapa model dasar. Dengan kata lain, mekanisme ensemble learning mampu mengurangi dampak ketidakseimbangan kelas tanpa memerlukan teknik oversampling tambahan

Secara keseluruhan, hasil pengujian statistik menunjukkan bahwa efektivitas SMOTE sangat bergantung pada jenis model yang digunakan. Pada model tunggal seperti EfficientNetB0, SMOTE memberikan kontribusi signifikan dalam meningkatkan performa, khususnya dalam mendeteksi kelas positif pneumonia. Namun, pada model Stacking Ensemble, pengaruh SMOTE tidak signifikan karena model tersebut sudah memiliki struktur pembelajaran yang lebih robust terhadap distribusi data yang tidak seimbang. Temuan ini menegaskan bahwa penggunaan SMOTE lebih relevan diterapkan pada model tunggal, sementara pendekatan ensemble learning dapat menjadi solusi yang lebih stabil untuk permasalahan klasifikasi medis dengan ketidakseimbangan kelas.

4.3.4 Implikasi Hasil Penelitian terhadap Aplikasi Klinis dan Integrasi Nilai Keislaman

Pengembangan sistem deteksi pneumonia berbasis stacking ensemble CNN tidak hanya memiliki manfaat secara ilmiah, tetapi juga selaras dengan nilai-nilai Islam terkait penjagaan jiwa (hifz an-nafs). Diagnosis yang akurat sangat penting untuk mencegah keterlambatan penanganan, sehingga teknologi yang meningkatkan sensitivitas dan spesifisitas model menjadi bentuk nyata dari upaya

menghindarkan bahaya (dar' al-mafāsīd). Hal ini sejalan dengan firman Allah SWT dalam QS. Al-Mā'idah ayat 32:

وَمَنْ أَحْيَاهَا فَكَأَنَّمَا أَحْيَا النَّاسَ جَمِيعًا

“Barang siapa memelihara kehidupan seorang manusia, maka seolah-olah ia telah memelihara kehidupan seluruh manusia.”

Ayat tersebut menegaskan bahwa menjaga keselamatan jiwa merupakan salah satu tujuan mulia yang sangat dijunjung tinggi. Dalam konteks penelitian ini, penyediaan sistem berbasis AI yang mampu meningkatkan deteksi dini pneumonia dapat mendukung tenaga medis dalam mengambil keputusan yang lebih cepat dan tepat, sehingga berpotensi menyelamatkan lebih banyak nyawa. Dengan demikian, inovasi teknologi kesehatan seperti ini dapat dipahami sebagai kontribusi terhadap kemaslahatan manusia dan selaras dengan nilai-nilai etis Islam.

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan rangkaian eksperimen yang dilakukan, dapat disimpulkan bahwa pendekatan stacking ensemble CNN memberikan performa paling unggul dan stabil dalam mendeteksi pneumonia pada citra X-ray paru-paru dibandingkan model tunggal EfficientNetB0, MobileNetV2, dan MobileNetV3Small. Model ini mencapai hasil terbaik pada skenario full dataset dengan rasio 80:20, ditandai akurasi 99,14%, sensitivitas 98,99%, dan AUC-ROC 99,87%, sekaligus menunjukkan sebaran metrik yang lebih rapat dan konsisten. Analisis juga mengungkap bahwa skenario subset—khususnya split 70:30—mengalami variabilitas performa lebih besar akibat keterbatasan jumlah data latih, sementara penggunaan SMOTE terbukti meningkatkan performa model tunggal secara signifikan namun hanya memberi dampak kecil pada stacking karena sifat ensemble yang memang lebih robust terhadap ketidakseimbangan kelas. Temuan ini menegaskan bahwa distribusi dan jumlah data sangat memengaruhi stabilitas model, dan bahwa stacking ensemble CNN merupakan pendekatan yang paling efektif, adaptif, serta layak diimplementasikan sebagai sistem pendukung skrining klinis yang membutuhkan sensitivitas tinggi untuk meminimalkan kasus pneumonia yang terlewat.

5.2 Saran

Penelitian ini membuka peluang untuk pengembangan lebih lanjut dalam penerapan *deep learning* pada bidang medis, khususnya untuk deteksi penyakit

paru-paru berbasis citra. Beberapa saran yang dapat diberikan untuk penelitian berikutnya adalah sebagai berikut. Pertama, penelitian selanjutnya disarankan menggunakan dataset yang lebih besar dan beragam, mencakup berbagai variasi demografis, kualitas citra, serta sumber rumah sakit yang berbeda. Pendekatan ini akan memperkuat kemampuan generalisasi model dan meningkatkan validitas eksternal hasil penelitian. Selain itu, penggunaan data multimodal seperti citra CT-scan atau rekam medis elektronik (EMR) dapat memperkaya konteks diagnosis.

Kedua, dari sisi implementasi, model stacking ensemble dapat dikembangkan menjadi prototipe *Clinical Decision Support System (CDSS)* yang terintegrasi dengan sistem rumah sakit. Uji coba langsung pada lingkungan klinis diperlukan untuk menilai keandalan model terhadap data nyata, termasuk pengujian waktu inferensi, konsumsi memori, dan akurasi diagnosis dalam kondisi *real-time*. Pendekatan *model compression* atau *quantization* juga dapat diterapkan agar sistem tetap efisien saat dijalankan pada perangkat dengan sumber daya terbatas.

Terakhir, dari perspektif akademik, penelitian lanjutan dapat diarahkan pada topik *Explainable AI (XAI)* dan *multi-modal biomedical learning*, yang menggabungkan citra medis dengan data laboratorium atau sensor fisiologis untuk meningkatkan konteks diagnostik. Selain itu, pendekatan optimasi hiperparameter otomatis seperti *Bayesian Optimization* atau *AutoML* dapat diterapkan untuk mempercepat proses eksperimental dan menemukan konfigurasi model yang optimal.

Dengan demikian, penelitian ini diharapkan dapat menjadi dasar bagi pengembangan sistem deteksi pneumonia berbasis AI yang tidak hanya akurat,

tetapi juga dapat dipertanggungjawabkan secara etis dan praktis dalam lingkungan klinis. Penerapan metode stacking ensemble pada data medis membuktikan bahwa kolaborasi antar model dapat menghasilkan solusi yang lebih andal, efisien, dan berdampak positif bagi diagnosis penyakit di masa depan.

DAFTAR PUSTAKA

- Al-Qur'an al-Karim. (2019). *Al-Qur'an dan Terjemahannya*. Jakarta: Kementerian Agama Republik Indonesia, Lajnah Pentashihan Mushaf Al-Qur'an.
- Ayan, E., & Ünver, H. M. (2019). Diagnosis of pneumonia from chest X-ray images using deep learning. *Procedia Computer Science*, 163, 85–92. <https://doi.org/10.1016/j.procs.2019.12.095>
- Cherian, T., Mulholland, E. K., Carlin, J. B., Ostensen, H., Amin, R., Campo, M., Greenberg, D., Lagos, R., Lucero, M., Madhi, S. A., O'Brien, K. L., Obaro, S., & Steinhoff, M. C. (2005). Standardized interpretation of paediatric chest radiographs for the diagnosis of pneumonia in epidemiological studies. *Bulletin of the World Health Organization*, 83(5), 353–359. <https://doi.org/10.1590/S0042-96862005000500011>
- Chouhan, V., Singh, S. K., Khamparia, A., Gupta, D., Tiwari, P., Moreira, C., Damaševičius, R., & de Albuquerque, V. H. C. (2020). A novel transfer learning-based approach for pneumonia detection in chest X-ray images. *Applied Sciences*, 10(2), 559. <https://doi.org/10.3390/app10020559>
- Dietterich, T. G. (2000). Ensemble methods in machine learning. *International Workshop on Multiple Classifier Systems*, 1–15.
- Hansell, D. M., Bankier, A. A., MacMahon, H., McLoud, T. C., Müller, N. L., & Remy, J. (2008). Fleischner Society: Glossary of terms for thoracic imaging. *Radiology*, 246(3), 697–722. <https://doi.org/10.1148/radiol.2462070712>
- Hashmi, M. F., Katiyar, S., Keskar, A. G., Bokde, N. D., & Geem, Z. W. (2020). Efficient pneumonia detection in chest X-ray images using deep transfer learning. *Diagnostics*, 10(6), 417. <https://doi.org/10.3390/diagnostics10060417>
- Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., ... Adam, H. (2019). Searching for MobileNetV3. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 1314–1324.
- Kallio, H., Kallio, M., Huttunen, J., Lehtimäki, T., & Lehtimäki, M. (2019). Artificial intelligence in radiology: Applications, challenges and opportunities. *Duodecim*, 135(19), 1951–1959.
- Kaggle Dataset — Mooney, P. (2018). *Chest X-Ray Images (Pneumonia)*. <https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia>
- Kermayn, D. S., Goldbaum, M., Cai, W., Valentim, C. C. S., Liang, H., Baxter, S. L., ... Zhang, K. (2018). Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5), 1122–1131.e9. <https://doi.org/10.1016/j.cell.2018.02.010>

- Lakhani, P., & Sundaram, B. (2017). Deep learning at chest radiography: Automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology*, 284(2), 574–582. <https://doi.org/10.1148/radiol.2017162326>
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444. <https://doi.org/10.1038/nature14539>
- Liang, G., & Yu, L. (2020). An ensemble model of deep learning for pneumonia detection in chest X-ray images. *Multimedia Tools and Applications*, 79(41), 30483–30496. <https://doi.org/10.1007/s11042-020-09067-0>
- Majkowska, A., Mittal, S., Steiner, D. F., Reicher, J. J., McKinney, S. M., Duggan, G. E., ... Rajpurkar, P. (2020). Chest radiograph interpretation with deep learning models: Assessment of radiologist-level accuracy for pneumonia detection. *Nature Medicine*, 26(6), 890–897. <https://doi.org/10.1038/s41591-020-0852-9>
- Mandell, L. A., Niederman, M. S., Aspa, J., Torres, A., & Welte, T. (2019). Infectious Diseases Society of America/American Thoracic Society consensus guidelines on the management of community-acquired pneumonia in adults. *Clinical Infectious Diseases*, 71(6), e136–e164. <https://doi.org/10.1093/cid/ciz1007>
- Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., ... Ng, A. Y. (2017). CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. *arXiv preprint arXiv:1711.05225*. <https://arxiv.org/abs/1711.05225>
- Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4510–4520.
- Sari, D. P., Utami, E., & Kusuma, G. P. (2021). Pneumonia classification on chest X-ray images using EfficientNet architecture. *Journal of Physics: Conference Series*, 1844(1), 012050. <https://doi.org/10.1088/1742-6596/1844/1/012050>
- Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 6105–6114.
- World Health Organization. (2020). *Pneumonia*. <https://www.who.int/news-room/fact-sheets/detail/pneumonia>
- Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS*

Medicine, 15(11), e1002683.
<https://doi.org/10.1371/journal.pmed.1002683>

- Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). *MobileNetV2: Inverted Residuals and Linear Bottlenecks*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 4510–4520.
- Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q. V., & Adam, H. (2019). *Searching for MobileNetV3*. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 1314–1324.
<https://doi.org/10.1109/ICCV.2019.00140>
- Tan, M., & Le, Q. V. (2019). *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*. In Proceedings of the 36th International Conference on Machine Learning (ICML), 6105–6114.
PMLR. <https://arxiv.org/abs/1905.11946>
- Almirall, J., Serra-Prat, M., Bolívar, I., & Balasso, V. (2017). Risk factors for community-acquired pneumonia in adults: A systematic review of observational studies. *The Lancet Respiratory Medicine*, 5(12), 1053–1065.
- Hansell, D. M., Bankier, A. A., MacMahon, H., McLoud, T. C., Müller, N. L., & Remy, J. (2008). Fleischner Society: Glossary of terms for thoracic imaging. *Radiology*, 246(3), 697–722.
- Majkowska, A., Mittal, S., Steiner, D. F., Reicher, J. J., McKinney, S. M., Duggan, G. E., ... & Corrado, G. S. (2020). Chest radiograph interpretation with deep learning models: Assessment using radiologist-adjudicated reference standards and population-adjusted evaluation. *Nature Medicine*, 26(7), 890–899.
- Mandell, L. A., Wunderink, R. G., Anzueto, A., Bartlett, J. G., Campbell, G. D., Dean, N. C., ... & Whitney, C. G. (2019). Infectious Diseases Society of America/American Thoracic Society consensus guidelines on the management of community-acquired pneumonia in adults. *Clinical Infectious Diseases*, 44(Supplement_2), S27–S72.
- Thorax. (2000). Guidelines for the management of community-acquired pneumonia in adults. *Thorax*, 55(4), 318–332.
- The Lancet. (2010). Global burden of disease from lower respiratory infections. *The Lancet*, 375(9727), 1545–1555.
- The Lancet. (2015). Ending preventable pneumonia deaths: What will it take? *The Lancet*, 386(9998), 107–108.

- ScienceDirect. (2010). Complications of severe pneumonia. *Respiratory Care Clinics*, 16(2), 243–260.
- PMC. (2020). Artificial intelligence in medical imaging: Pneumonia detection applications. *Journal of Medical Systems*, 44(7), 1–12.