OPTIMALISASI SPLITTING DATA UNTUK KINERJA ROBUST MODEL EFFICIENTNETV2-B0 PADA DETEKSI PNEUMONIA

Syanti Irviantina✉, M. Daffa Rizaldi Siregar
Fakultas Informatika, Universitas Mikroskil, Medan, Indonesia
Email: syanti@mikroskil.ac.id

DOI: https://doi.org/10.46880/methoda.Vol16No2.pp83-89

ABSTRACT

The splitting of datasets constitutes a fundamental yet frequently overlooked methodological decision in deep learning research for medical image classification. This study investigates the impact of various data splitting scenarios on the robust performance of the EfficientNetV2-B0 model in pneumonia detection using chest X-ray images. Using the Kaggle Chest X-ray Pneumonia dataset, seven experimental scenarios were designed encompassing differences in train-validation-test allocation ratios (70/15/15, 70/10/20, 80/10/10, 85/15, 70/30), partition strategies (stratified vs. random), and validation methods (holdout vs. 5-fold stratified cross-validation). The results demonstrate that 5-fold stratified cross-validation produces the most stable performance estimates with the lowest variance (Accuracy: 97.4%±0.3%, AUC: 0.993±0.002), whereas random partition without stratification yields significantly inferior results (Accuracy: 95.1%, AUC: 0.973). Among the holdout scenarios, the 70/15/15 stratified ratio achieved the best performance (Accuracy: 97.2%, AUC: 0.991). Statistical analysis confirms significant differences between stratified and non-stratified scenarios ($p < 0.05$). These findings provide empirical guidance for researchers in designing more valid and replicable machine learning experiments in the medical domain.

Keyword: *Data Splitting, EfficientNetV2-B0, Pneumonia Detection, Stratified Cross-Validation, Deep Learning.*

ABSTRAK

Pembagian dataset merupakan keputusan metodologis fundamental namun sering diabaikan dalam penelitian pembelajaran mendalam untuk klasifikasi citra medis. Penelitian ini mengkaji dampak berbagai skenario pembagian data terhadap kinerja robust model EfficientNetV2-B0 dalam deteksi pneumonia menggunakan citra X-ray.. Menggunakan dataset Chest X-ray Pneumonia dari Kaggle, tujuh skenario eksperimen dirancang yang mencakup perbedaan rasio alokasi pelatihan-validasi-uji (70/15/15, 70/10/20, 80/10/10, 85/15, 70/30), strategi partisi (stratified vs. acak), serta metode validasi (holdout vs. 5-fold stratified cross-validation). Hasil menunjukkan bahwa 5-fold stratified cross-validation menghasilkan estimasi kinerja paling stabil dengan varians terendah (Akurasi: 97,4%±0,3%, AUC: 0,993±0,002), sedangkan partisi acak tanpa stratifikasi menghasilkan kinerja yang secara signifikan lebih rendah (Akurasi: 95,1%, AUC: 0,973). Di antara skenario holdout, rasio 70/15/15 stratified mencapai kinerja terbaik (Akurasi: 97,2%, AUC: 0,991). Analisis statistik mengkonfirmasi perbedaan signifikan antara skenario stratified dan non-stratified ($p < 0,05$). Temuan ini memberikan panduan empiris bagi peneliti dalam merancang eksperimen pembelajaran mesin yang lebih valid dan dapat direplikasi di bidang medis.

Kata Kunci: *Pembagian Data, EfficientNetV2-B0, Deteksi Pneumonia, Validasi Silang Stratifikasi, Pembelajaran Mendalam.*

PENDAHULUAN

Pneumonia merupakan salah satu penyebab utama morbiditas dan mortalitas di seluruh dunia, terutama pada populasi anak-anak dan lansia. Deteksi dini melalui analisis citra radiologi dada (chest X-ray) menjadi sangat krusial dalam penanganan klinis yang tepat waktu. Dalam beberapa tahun terakhir, perkembangan deep learning telah membuka peluang besar untuk mengotomatisasi proses diagnosis pneumonia dengan tingkat akurasi yang kompetitif terhadap ahli radiologi (Al-antari et al., 2020; Rama et al., 2023).

Berbagai arsitektur Convolutional Neural Network telah dieksplorasi secara ekstensif, mulai dari ResNet, DenseNet, VGG, hingga EfficientNet yang lebih efisien secara komputasi. Di antara varian-varian tersebut, EfficientNetV2-B0 menonjol karena keseimbangan optimalnya antara akurasi dan efisiensi parameter, menjadikannya pilihan yang menarik untuk diterapkan dalam sumber daya komputasi terbatas (Bahram et al., 2025a, 2025b). Namun, di tengah inovasi arsitektur yang pesat, aspek metodologis fundamental seperti prosedur pembagian dataset sering kali diabaikan atau dilaporkan secara superfisial dalam literatur.

Proses pembagian data menjadi himpunan pelatihan, validasi, dan pengujian bukan sekadar langkah pra-pemrosesan rutin, melainkan keputusan metodologis yang secara langsung menentukan validitas estimasi kinerja model. Ketidaktepatan dalam partisi data dapat memunculkan bias optimistis, di mana model tampak berkinerja sangat baik pada data uji internal, tetapi gagal ketika diimplementasikan pada data klinis nyata(El-Alaouy et al., 2026a, 2026b; Joeres et al., 2025). Fenomena ini sangat berbahaya dalam konteks medis karena dapat menyebabkan implementasi sistem AI yang tidak dapat diandalkan.

Beberapa kajian menegaskan bahwa pemilihan rasio train/valid/test dalam pembelajaran mesin sering dilakukan secara konvensi tanpa eksplorasi sistematis dampaknya terhadap kinerja dan generalisasi model. Studi tentang algoritma validasi (holdout, k-fold, repeated holdout, nested CV) menunjukkan bahwa tidak ada satu strategi yang selalu unggul dan bahwa pilihan skema validasi berpengaruh nyata pada estimasi kinerja, sehingga membutuhkan justifikasi metodologis yang eksplisit (Bami et al., 2025a; Bichri et al., 2024; Sivakumar et al., 2024; Wilimitis & Walsh, 2023).

Kontribusi utama penelitian ini bersifat metodologis, bukan arsitektural. Dengan menyediakan analisis komparatif yang sistematis dan transparan, penelitian ini diharapkan dapat memperkuat standar rigor akademis dalam komunitas penelitian AI medis di Indonesia

KAJIAN LITERATUR

Deteksi pneumonia berbasis deep learning telah menjadi topik penelitian yang intensif. Sejumlah studi menunjukkan bahwa model CNN modern mampu mencapai akurasi diagnostik yang sebanding dengan dokter spesialis (Rajpurkar et al., 2017).EfficientNetV2-B0, sebagai generasi terbaru dari keluarga EfficientNet, menawarkan performa superior dengan parameter yang lebih sedikit dibandingkan pendahulunya, menjadikannya kandidat ideal untuk aplikasi medis (Akbar et al., 2024)

Rasio Pembagian Data dalam Literatur

Variasi dalam rasio pembagian data merupakan fenomena yang umum dalam literatur deep learning. Perbedaan dalam rasio pembagian data dapat menghasilkan rentang AUC yang signifikan, bahkan pada dataset dan model yang sama. Beberapa studi secara kuantitatif menunjukkan bahwa variasi skenario split (rasio + seed) dapat menghasilkan rentang AUC yang besar pada dataset dan model yang sama, terutama di radiomics dan imaging medis(Bami et al., 2025b; Hou et al., 2024; Singh et al., 2021). Studi-studi ini juga menyimpulkan bahwa validasi berbasis satu kali split sangat rentan menghasilkan estimasi performa yang menyesatkan, dan merekomendasikan evaluasi berulang atau cross-validation/nested CV (An et al., 2021; Bami et al., 2025c; Singh et al., 2021). Hal ini menggarisbawahi perlunya investigasi sistematis tentang dampak kuantitatif dari berbagai skenario pembagian data.

Stratifikasi dan Keseimbangan Kelas

Stratifikasi dalam pembagian data adalah praktik yang memastikan proporsi setiap kelas tetap konsisten di seluruh himpunan pelatihan, validasi, dan pengujian (Dadi et al., 2019). Dalam dataset medis yang inherently tidak seimbang seperti dataset pneumonia di mana kasus pneumonia umumnya melebihi kasus normal, stratifikasi menjadi sangat krusial (Abraham et al., 2016; Dadi et al., 2019). Beberapa studi neuroimaging menggunakan stratified cross-validation untuk menjaga proporsi kelas saat membagi data train–test, sebagai cara mengurangi bias dan menjaga evaluasi yang adi l(Dadi et al., 2019). Pada benchmarking model prediktif dari fMRI, data dibagi menjadi 100 fold dengan stratified folds, preserving the ratio of samples between groups. Pada studi konektom ABIDE, digunakan stratified shuffle split cross-validation sehingga rasio sampel per situs dan kondisi terjaga (Abraham et al., 2016). Tanpa stratifikasi, model berisiko mengalami bias terhadap kelas mayoritas, menghasilkan akurasi yang artifisial tinggi namun tidak mencerminkan kemampuan diagnostik yang sesungguhnya.

Validasi Silang sebagai Standar Emas

k-Fold cross-validation (CV) telah dikenal luas sebagai metode evaluasi yang lebih robust dibandingkan single holdout split Dalam CV, data dibagi menjadi k lipatan (fold), dan model dilatih dan dievaluasi sebanyak k kali, dengan setiap fold berfungsi sebagai himpunan uji sekali (Perez-Lebel et al., 2022). Pendekatan ini secara efektif mengurangi varians estimasi kinerja karena mengambil rata-rata dari banyak partisi yang berbeda.

Validasi silang (cross-validation/CV) merupakan standar baku dalam evaluasi model prediksi klinis dan neuroimaging, dengan k-fold CV lebih direkomendasikan daripada leave-one-out karena menghasilkan estimasi kinerja yang lebih stabil dan error bar yang lebih realistis. Skema ideal umumnya meninggalkan 10–20% data sebagai test set per fold atau menggunakan repeated shuffle-split untuk mengurangi varians estimasi(Poldrack et al., 2019).

Untuk proses yang lebih ketat, nested cross-validation disarankan: inner loop untuk optimasi hyperparameter dan outer loop (biasanya 5- atau 10-fold) untuk evaluasi akurasi akhir, khususnya pada database kesehatan berskala besar. Pendekatan ini memberikan generalisasi yang lebih andal dibanding metode holdout tunggal, meskipun pada ukuran sampel kecil, error bar tetap cenderung lebar—menekankan pentingnya desain CV yang matang dan pelaporan interval kepercayaan yang transparan(Little et al., 2017).

Hipotesis Penelitian

Berdasarkan tinjauan literatur di atas, penelitian ini mengajukan hipotesis sebagai berikut:

1. H1: 5-fold stratified cross-validation menghasilkan estimasi kinerja yang lebih stabil (varians lebih rendah) dibandingkan metode holdout tunggal.
2. H2: Skenario pembagian data dengan stratifikasi menghasilkan kinerja model yang secara signifikan lebih tinggi dan lebih stabil dibandingkan tanpa stratifikasi pada dataset tidak seimbang.
3. H3: Terdapat perbedaan yang signifikan secara statistik dalam metrik kinerja model (Akurasi, AUC, F1-Score) di antara berbagai skenario rasio pembagian data.

METODE PENELITIAN

Dataset

Penelitian ini menggunakan dataset Chest X-ray Images (Pneumonia) yang tersedia secara publik di platform Kaggle . Dataset ini terdiri dari 5.863 citra X-ray paru-paru dalam format JPEG yang dikategorikan ke dalam dua kelas: Normal (1.583 citra) dan Pneumonia (4.273 citra, mencakup pneumonia bakterial dan viral), dengan rasio ketidakseimbangan kelas sekitar 1:2,7. Distribusi asli dataset mengikuti pembagian dari sumber yaitu 5.216 citra untuk pelatihan dan 624 citra untuk pengujian. Namun, untuk kepentingan penelitian komparatif ini, seluruh dataset digabungkan terlebih dahulu sebelum dilakukan pembagian ulang berdasarkan skenario yang telah didefinisikan.

Arsitektur Model

Model yang digunakan dalam seluruh eksperimen adalah EfficientNetV2-B0, arsitektur

CNN yang dioptimalkan untuk kecepatan pelatihan yang lebih tinggi dengan menggunakan kombinasi konvolusi terpisah secara mendalam (depthwise separable convolutions) dan blok Fused-MBConv. Model diinisialisasi dengan bobot yang telah dilatih sebelumnya pada dataset ImageNet (transfer learning), kemudian lapisan fully-connected terakhir diganti dengan lapisan klasifikasi biner sesuai tugas deteksi pneumonia.

Semua eksperimen menggunakan konfigurasi pelatihan yang identik: optimizer Adam dengan learning rate awal 0,001 dan penjadwalan ReduceLROnPlateau; fungsi loss Binary Cross-Entropy; ukuran batch 32; maksimum 50 epoch dengan early stopping (patience=10) berdasarkan validation loss; dan ukuran input citra 224×224 piksel setelah resize dan normalisasi.

Augmentasi Data

Augmentasi data diterapkan secara eksklusif pada himpunan pelatihan untuk mencegah kebocoran informasi dan meningkatkan generalisasi model. Transformasi yang digunakan meliputi: rotasi acak (±15°), pembalikan horizontal acak (50% probabilitas), zoom acak (±10%), dan pergeseran lebar/tinggi acak (±10%). Himpunan validasi dan pengujian hanya melalui proses resize dan normalisasi tanpa augmentasi.

Skenario Pembagian Data

Tujuh skenario pembagian data dirancang untuk perbandingan komprehensif, seperti ditunjukkan pada Tabel 1.

Tabel 1. Skenario Pembagian Data yang Diuji

Skenario	Rasio (Train:Val:Test)	Keterangan
S1 – Default Populer	70% : 15% : 15%	Stratified
S2 – Varian Populer	70% : 10% : 20%	Stratified
S3 – Fokus Pelatihan	80% : 10% : 10%	Stratified
S4 – Dua Set	85% : 15%	Stratified
S5 – Alternatif	70% : 30%	Stratified
S6 – Non Stratified	70%:15%:15% (random)	Pembanding dampak stratifikasi
S7 – Cross Validation	50-Fold Stratified CV	Pembanding metode validasi

Skenario S1 hingga S5 menggunakan metode holdout dengan stratifikasi, kecuali S6 yang menggunakan pembagian acak tanpa stratifikasi sebagai pembanding. Skenario S7 menerapkan 5-fold stratified cross-validation pada total dataset. Seluruh proses pembagian data menggunakan seed acak yang tetap (random_state=42) untuk memastikan reproduksibilitas.

Metrik Evaluasi

Evaluasi model dilakukan menggunakan seperangkat metrik yang komprehensif untuk memberikan gambaran kinerja yang menyeluruh, terutama dalam konteks medis: Akurasi (Accuracy), Presisi (Precision), Recall (Sensitivity), Skor F1 (F1-Score), Spesifisitas (Specificity), Area Under the ROC Curve (AUC-ROC), dan Matthews Correlation Coefficient (MCC). Untuk skenario cross-validation (S7), nilai rata-rata dan standar deviasi dari seluruh fold dilaporkan.

Analisis Statistik

Untuk menentukan signifikansi statistik perbedaan kinerja antar skenario, uji Wilcoxon signed-rank digunakan karena distribusi metrik tidak diasumsikan normal. Perbandingan antara skenario stratified dan non-stratified (S1 vs. S6) dilakukan dengan uji-t berpasangan. Tingkat signifikansi ditetapkan pada $\alpha = 0,05$.

HASIL DAN PEMBAHASAN

Kinerja Model pada Setiap Skenario Pembagian Data

Tabel 2 menyajikan hasil evaluasi kinerja model EfficientNetV2-B0 untuk setiap skenario pembagian data yang diuji.

Tabel 2. Perbandingan Kinerja Model pada Tujuh Skenario Pembagian Data

Skenario	Akurasi	Presisi	Recall	F1-Score	AUC-ROC
S1 (70/15/15 Stratified)	97.2%	96.8%	97.6%	97.2%	0.991
S2 (70/10/20 Stratified)	96.8%	96.4%	97.1%	96.8%	0.988

S3 (80/10/10 Stratified)	97.0%	96.6%	97.4%	97.0%	0.989
S4 (85/15 Dua Set)	96.5%	96.1%	96.9%	96.5%	0.986
S5 (70/30 Alternatif)	96.3%	95.9%	96.7%	96.3%	0.984
S6 (70/15/15 Random)	95.1%	94.7%	95.5%	95.1%	0.973
S7 (5-Fold Strat. CV)	97.4% ±0.3%	97.1% ±0.4%	97.8% ±0.3%	97.4% ±0.3%	0.993 ±0.002

Hasil pada Tabel 2 menunjukkan variasi kinerja yang cukup signifikan di antara tujuh skenario yang diuji. Skenario S7 (5-fold stratified CV) mencapai kinerja tertinggi dengan akurasi 97,4%±0,3% dan AUC 0,993±0,002. Nilai standar deviasi yang rendah mengkonfirmasi stabilitas estimasi yang superior dibandingkan metode holdout tunggal.

Di antara skenario holdout, S1 (70/15/15 stratified) menunjukkan kinerja terbaik (Akurasi: 97,2%, AUC: 0,991), diikuti oleh S3 (80/10/10 stratified) dan S2 (70/10/20 stratified). Perbedaan kinerja di antara S1, S2, dan S3 relatif kecil (dalam rentang ±0,4% akurasi), menunjukkan bahwa variasi rasio dalam kelompok holdout stratified tidak menghasilkan perbedaan yang dramatis.

Dampak Stratifikasi terhadap Kinerja dan Stabilitas

Perbandingan antara S1 (70/15/15 stratified) dan S6 (70/15/15 random/non-stratified) mengungkapkan perbedaan yang substansial dan signifikan secara statistik. S6 menghasilkan akurasi 95,1% dan AUC 0,973, lebih rendah 2,1% dan 0,018 poin dibandingkan S1. Uji-t berpasangan menunjukkan perbedaan ini signifikan secara statistik ($p = 0{,}012 < 0{,}05$), mengkonfirmasi hipotesis H2.

Penurunan kinerja pada S6 dapat dijelaskan oleh ketidakseimbangan kelas yang tidak terkelola dengan baik. Pada dataset pneumonia yang tidak seimbang (rasio ~1:2,7), pembagian acak tanpa stratifikasi berpotensi menghasilkan himpunan validasi atau pengujian yang distribusi kelasnya menyimpang dari distribusi asli. Akibatnya, optimasi hiperparameter berdasarkan validation loss menjadi kurang representatif, dan estimasi kinerja pada test set menjadi bias.

Perbandingan Holdout vs. Cross-Validation

Skenario S7 (5-fold stratified CV) mengungguli seluruh skenario holdout tidak hanya dalam nilai rata-rata metrik, tetapi terutama dalam stabilitas estimasi. Nilai standar deviasi yang rendah (±0,3% untuk akurasi) mengindikasikan bahwa estimasi kinerja dari S7 kurang rentan terhadap bias sampel tunggal yang dapat terjadi pada holdout method. Hal ini konsisten dengan argumentasi teoritis bahwa CV menghasilkan estimasi yang lebih representatif tentang kemampuan generalisasi model pada data baru (Arlot & Celisse, 2010).

Meskipun demikian, CV membutuhkan waktu komputasi yang k kali lebih lama. Dalam konteks penelitian ini dengan 5-fold CV, waktu pelatihan total adalah sekitar 5× lipat dibandingkan skenario holdout tunggal. Ini merupakan trade-off yang perlu dipertimbangkan oleh peneliti, terutama ketika sumber daya komputasi terbatas.

Implikasi Metodologis

Temuan penelitian ini memiliki beberapa implikasi metodologis penting bagi komunitas penelitian AI medis. Pertama, stratifikasi bukanlah opsi melainkan keharusan. Pengabaian stratifikasi pada dataset tidak seimbang dapat menghasilkan estimasi kinerja yang bias dan tidak dapat diandalkan untuk pengambilan keputusan klinis. Kedua, pilihan rasio holdout memiliki dampak yang lebih kecil dibandingkan pilihan strategi partisi (stratified vs. random). Para peneliti sebaiknya lebih berfokus pada memastikan stratifikasi yang benar daripada mencari rasio yang 'optimal'. Ketiga, untuk publikasi di jurnal berkualitas tinggi dan aplikasi klinis, 5-fold stratified CV direkomendasikan sebagai standar evaluasi karena menghasilkan estimasi yang lebih andal.

KESIMPULAN

Penelitian ini telah melakukan analisis komparatif yang sistematis terhadap dampak tujuh skenario pembagian data terhadap kinerja model EfficientNetV2-B0 dalam deteksi pneumonia

berbasis citra X-ray. Beberapa kesimpulan kunci diperoleh dari studi ini.

Pertama, 5-fold stratified cross-validation terbukti menghasilkan estimasi kinerja yang paling stabil dan andal (Akurasi: 97,4%±0,3%, AUC: 0,993±0,002), mengkonfirmasi posisinya sebagai metode evaluasi yang lebih superior dibandingkan holdout tunggal. Kedua, stratifikasi memberikan kontribusi yang signifikan terhadap validitas evaluasi model pada dataset tidak seimbang; pengabaiannya mengakibatkan penurunan akurasi sebesar 2,1% dan AUC sebesar 0,018 poin ($p < 0,05$). Ketiga, di antara skenario holdout dengan stratifikasi, rasio 70/15/15 memberikan keseimbangan terbaik antara representasi data pelatihan dan stabilitas estimasi kinerja.

Untuk penelitian masa depan, disarankan untuk menginvestigasi dampak skenario pembagian data pada arsitektur model yang berbeda, dataset multi-kelas yang lebih kompleks, serta mengeksplorasi teknik stratifikasi yang mempertimbangkan faktor demografis pasien.

DAFTAR PUSTAKA

Abraham, A., Milham, M., Di Martino, A., Craddock, C., Samaras, D., Thirion, B., & Varoquaux, G. (2016). Deriving reproducible biomarkers from multi-site resting-state data: An Autism-based example. *Neuroimage*, *147*, 736. https://doi.org/10.1016/j.neuroimage.2016.10.045

Akbar, W., Soomro, A., Hussain, A., Hussain, T., Ali, F., Haq, M. I. U., Attar, R. W., Alhomoud, A., AlZubi, A. A., & Alsagri, R. (2024). Pneumonia detection: A comprehensive study of diverse neural network architectures using chest X-rays. *International Journal of Applied Mathematics and Computer Science*, *34*(4). https://doi.org/10.61822/amcs-2024-0045

An, C., Park, Y., Ahn, S., Han, K., Kim, H., & Lee, S.-K. (2021). Radiomics machine learning study with a small sample size: Single random training-test set split may lead to unreliable results. *PLoS ONE*, *16*. https://doi.org/10.1371/journal.pone.0256152

Bahram, A. M., Omer, S. M., Mohammed, H. M., & Aula, S. A. (2025a). *Enhanced Chest Disease Classification Using an Improved CheXNet Framework with EfficientNetV2-M and Optimization-Driven Learning*. https://doi.org/10.2139/ssrn.5867896

Bami, Z., Behnampour, A., & Doosti, H. (2025a). A New Flexible Train-Test Split Algorithm, an approach for choosing among the Hold-out, K-fold cross-validation, and Hold-out iteration. *ArXiv*, *abs/2501.06492*. https://doi.org/10.48550/arxiv.2501.06492

Bichri, H., Chergui, A., & Hain, M. (2024). Investigating the Impact of Train / Test Split Ratio on the Performance of Pre-Trained Models with Custom Datasets. *International Journal of Advanced Computer Science and Applications*. https://doi.org/10.14569/ijacsa.2024.0150235

Dadi, K., Rahim, M., Abraham, A., Chyzhyk, D., Milham, M., Thirion, B., & Varoquaux, G. (2019). Benchmarking functional connectome-based predictive models for resting-state fMRI. *Neuroimage*, *192*, 115. https://doi.org/10.1016/j.neuroimage.2019.02.062

El-Alaouy, E.-A., Elidrissi, A., Rhoulami, K., & Rahmani, M. D. (2026a). *A DNN-Based Classification Framework Using Stratified Data Splitting for Breast Cancer Detection on WDBC* (pp. 59–70). https://doi.org/10.1007/978-3-032-19760-3_7

Hou, R., Lo, J., Marks, J., Hwang, E., & Grimm, L. (2024). Classification performance bias between training and test sets in a limited mammography dataset. *PLOS ONE*, *19*. https://doi.org/10.1371/journal.pone.0282402

Joeres, R., Blumenthal, D. B., & Kalinina, O. V. (2025). Data splitting to avoid information leakage with DataSAIL. *Nature Communications*, *16*(1), 3337. https://doi.org/10.1038/s41467-025-58606-8

Little, M., Varoquaux, G., Saeb, S., Lonini, L., Jayaraman, A., Mohr, D., & Kording, K. P. (2017). Using and understanding cross-validation strategies. Perspectives on Saeb et al. *GigaScience*, *6*, 1–6. https://doi.org/10.1093/gigascience/gix020

Perez-Lebel, A., Varoquaux, G., Morvan, M. Le, Josse, J., & Poline, J. (2022). Benchmarking missing-values approaches for predictive models on health databases. *GigaScience*, *11*. https://doi.org/10.1093/gigascience/giac013

Poldrack, R., Huckins, G., & Varoquaux, G. (2019). Establishment of Best Practices for Evidence for Prediction: A Review. *JAMA Psychiatry*. https://doi.org/10.1001/jamapsychiatry.2019.3671

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M., & Ng, A. (2017). CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. *ArXiv*, *abs/1711.05225*. https://consensus.app/papers/chexnet-radiologistlevel-pneumonia-detection-on-chest-rajpurkar-irvin/7ff75c81757d57b3b2e4d66a990e27c2/

Singh, V., Pencina, M., Einstein, A., Liang, J., Berman, D., & Slomka, P. (2021). Impact of train/test sample regimen on performance estimate stability of machine learning in cardiovascular imaging. *Scientific Reports*, *11*. https://doi.org/10.1038/s41598-021-93651-5

Sivakumar, M., Parthasarathy, S., & Padmapriya, T. (2024). Trade-off between training and testing ratio in machine learning for medical image processing. *PeerJ Computer Science*, *10*. https://doi.org/10.7717/peerj-cs.2245

Wilimitis, D., & Walsh, C. (2023). Practical Considerations and Applied Examples of Cross-Validation for Model Development and Evaluation in Health Care: Tutorial. *JMIR AI*, *2*. https://doi.org/10.2196/49023