Journal of Artificial Intelligence and Software Engineering
Vol. 5, No. 2, Juni 2025, pp. 697~708
E-ISSN: 2777-001X, DOI: 10.30811/jaise.v5i2.7043

Interpreting Lung Disease Detection from Chest *X-ray*s Using Layer-wise Relevance Propagation (LRP)

Laila Nurul Fauziyyah[1], Benny Sukma Negara[2], Muhammad Irsyad[3], Iwan Iskandar[4], Febi Yanto[5]
[1,2,3,4,5] Teknik Informatika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Sultan Syarif Kasim Riau, Pekanbaru, 28293, Indonesia'

Informasi Artikel

Diterima : 30 Mei 2025
Revisi : 9 Juni 2025
Publikasi : 20 Juni 2025

Kata Kunci:

Klasifikasi Citra
X-ray dada
VGG16
Interpretabilitas
Layer-wise Relevance Propagation

ABSTRAK

Penelitian ini mengusulkan pendekatan klasifikasi penyakit paru berbasis citra X-ray menggunakan arsitektur VGG16 yang dilengkapi metode interpretabilitas Layer-wise Relevance Propagation (LRP). Dataset terdiri dari tiga kelas: COVID-19, pneumonia, dan normal, yang diproses melalui augmentasi dan normalisasi. Model dilatih dengan rasio data 70:30, learning rate 0.001, batch size 32, dan optimizer Adam. Hasil pelatihan menunjukkan akurasi tinggi sebesar 96,78% dengan nilai precision, recall, dan F1-score yang seimbang. Metode LRP digunakan untuk menyoroti area penting pada citra yang berkontribusi terhadap prediksi model, sehingga meningkatkan transparansi keputusan. Kontribusi utama penelitian ini adalah integrasi VGG16 dengan LRP dalam klasifikasi multi-kelas citra X-ray, yang memberikan hasil akurat sekaligus interpretasi visual yang mendukung kepercayaan dalam aplikasi medis.

ABSTRACT

This study proposes an approach to classify lung diseases based on X-ray images using the VGG16 architecture equipped with the Layer-wise Relevance Propagation (LRP) interpretability method. The dataset consists of three classes: COVID-19, pneumonia, and normal, which are processed through augmentation and normalization. The model is trained with a data ratio of 70:30, a learning rate of 0.001, a batch size of 32, and an Adam optimizer. The training results show high accuracy of 96.78% with balanced precision, recall, and F1-score values. The LRP method was used to highlight important areas in the image that contributed to the model's prediction, thereby increasing decision transparency. The main contribution of this research is the integration of VGG16 with LRP in multi-class X-ray image classification, which provides accurate results along with visual interpretations that support confidence in medical applications.

***Penulis Koresponden**

Email: bsnegara@uin-suska.ac.id

L.N. Fauziyyah, B.S. Negara, M. Irsyad, I. Iskandar, & F. Yanto "Interpreting Lung Disease Detection from Chest X-rays Using Layer-wise Relevance Propagation (LRP)" *Journal of Artificial Intelligence and Software Engineering (J-AISE)*, vol. 5, no. 2, pp. 697-708, Juni 2025. doi: 10.30811/jaise.v5i2.7043

1. PENDAHULUAN

Penyebab utama kematian di seluruh dunia adalah penyakit paru-paru [1]. Berbagai macam penyakit yang mempengaruhi fungsi pernapasan dan kualitas hidup manusia, diantaranya adalah *Coronavirus Disease*

2019 (COVID-19) dan Pneumonia. COVID-19 merupakan penyakit yang dapat ditularkan melalui kontak langsung dengan penderita yang dihasilkan ketika orang yang terinfeksi batuk, bersin, atau menghembuskan napas. Seseorang dapat terinfeksi melalui pernapasan jika mereka berada dalam jarak dekat dengan seseorang yang menderita COVID-19 [2]. Sementara Pneumonia menjadi penyumbang kematian nomor 1 di Indonesia pada kelomopok *post neonatal* (usia 29 hari - 11 bulan) yaitu 14% naik dari tahun 2020 yaitu 9,8% kematian [3]. Resiko penularan COVID-19 yang tinggi dan angka kematian yang besar akibat Pneumonia menunjukkan diagnosis yang penting dan akurat terhadap penyakit paru-paru. Diagnosis penyakit paru-paru dapat dilakukan dengan anamnesa, pemeriksaan fisik, analisis sputum atau dahak, serta pemeriksaan laboratorium tambahan. Salah satu metode dalam pemeriksaan laboratorium tambahan adalah dengan menggunakan sinar-X (*X-ray*). Pemeriksaan *X-ray* dada (*Chest X-ray*) menjadi pilihan praktis untuk *screening* awal karena ketersediaan alat yang mudah didapatkan [4]. *Chest X-ray* (CXR) memvisualisasikan struktur paru-paru yang digunakan untuk mendeteksi kelainan seperti nodul, massa, dan akumulasi cairan, yang merupakan tanda-tanda penyakit paru-paru [5].

Disisi lain, kemajuan bidang teknologi, khususnya *deep learning* mendorong pengembangan model yang efektif untuk mendeteksi penyakit berbasis citra medis [6]. Arsitektur yang paling banyak digunakan adalah *convolutional neural network* (CNN) yang mampu mengenali pola-pola kompleks dalam gambar, termasuk citra *X-ray* yang sering kali tidak dapat dideteksi oleh mata manusia. Sejumlah penelitian telah menunjukkan bahwa CNN memiliki performa yang andal dalam praktik klinis untuk tugas klasifikasi medis dan telah digunakan secara luas. Penggunaan CNN berhasil mencapai tingkat akurasi hingga 99,90% dalam tugas klasifikasi penyakit [7]. Salah satu arsitektur CNN yang banyak digunakan adalah VGG16, yang dikenal efektif dalam ekstraksi fitur visual. Studi [8] melaporkan bahwa penerapan *transfer learning* menggunakan VGG16 untuk deteksi tumor otak pada citra MRI berhasil mencapai akurasi hingga 100% yang menunjukkan potensi besar pendekatan ini dalam aplikasi klinis. Implementasi model CNN dalam mendeteksi penyakit menggunakan citra berpotensi meningkatkan kualitas diagnosis medis.

Kelemahan utama dari *deep learning* adalah perilakunya sebagai *black-box*, sehingga sulit dipahami bagaimana model mengambil keputusan [9]. Hal ini mendorong berkembangnya bidang *explainable artificial intelligence* (XAI) yang bertujuan menjelaskan dan memberikan pemahaman yang logic di balik keputusan AI dan mengoptimalkan algoritma menjadi lebih baik. Penggunaan XAI dalam bidang kesehatan, penting untuk memberikan interpretabilitas hasil keputusan model sehingga meningkatkan kepercayaan penggunaan sistem dalam diagnosa [10]. Salah satu metode XAI yang dapat digunakan adalah *layer-wise relevance propagation* (LRP).

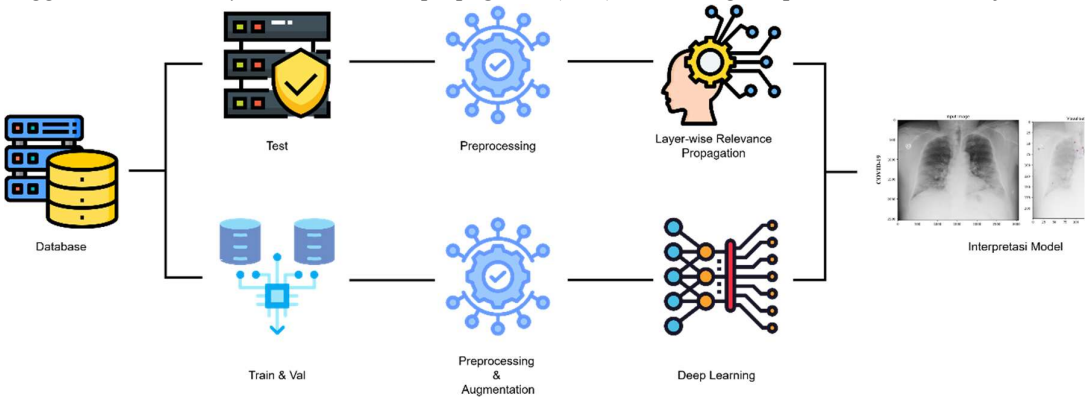
Layer-wise Relevance Propagation (LRP) adalah salah satu teknik yang dirancang untuk memberikan penjelasan tentang keputusan model *deep learning*, dengan cara memetakan kembali relevansi prediksi model ke bagian tertentu dari *input*, seperti area spesifik pada citra *X-ray* [11]. Metode ini memungkinkan untuk memahami dengan lebih jelas fitur atau area mana yang paling berpengaruh terhadap hasil yang diberikan oleh model [12]. Dalam konteks medis, LRP membantu mengidentifikasi bagian-bagian pada gambar yang paling berkontribusi terhadap keputusan model, sehingga hasil prediksi menjadi lebih transparan dan mudah dipahami oleh tenaga medis [13]. Salah satu penerapan LRP yang telah terbukti efektif ditunjukkan dalam penelitian sebelumnya yang mengintegrasikan LRP ke dalam model VGG16 untuk tugas klasifikasi citra sinar-X ginjal-ureter-kandung kemih (KUB). Studi tersebut berhasil mengidentifikasi keberadaan batu ginjal dengan akurasi tinggi, sekaligus memberikan visualisasi area citra yang menjadi fokus perhatian model melalui pendekatan *explainable artificial intelligence* (XAI) [14]. Kombinasi antara *transfer learning* dan LRP terbukti mampu menghasilkan interpretabilitas yang kuat serta meningkatkan kepercayaan pengguna terhadap hasil prediksi.

Meskipun demikian, cakupan penelitian tersebut masih terbatas pada klasifikasi biner (batu ginjal vs. normal) dan hanya diterapkan pada satu jenis citra medis, yaitu *X-ray* KUB. Belum banyak eksplorasi dilakukan pada penerapan LRP dalam klasifikasi citra medis lain yang lebih kompleks, seperti *X-ray* dada yang mencakup berbagai kondisi penyakit dengan karakteristik visual yang saling tumpang tindih. Selain itu, tantangan dalam klasifikasi multi-kelas serta kebutuhan akan interpretabilitas yang tinggi dalam pengambilan keputusan medis menjadi celah yang masih belum banyak dijembatani.

Penelitian ini hadir untuk menjawab keterbatasan tersebut dengan menerapkan pendekatan serupa dalam konteks yang lebih luas, yaitu klasifikasi citra *X-ray* dada ke dalam tiga kategori penyakit paru: COVID-19, paru-paru normal, dan pneumonia. Dalam konteks ini, interpretabilitas model tidak hanya penting untuk memahami keputusan algoritma, tetapi juga sangat krusial dalam meningkatkan kepercayaan klinis terhadap sistem deep learning. Oleh karena itu, studi ini bertujuan tidak hanya untuk mencapai akurasi klasifikasi yang tinggi melalui model VGG16, tetapi juga untuk menghasilkan interpretasi visual yang jelas dan bermakna menggunakan metode LRP, sehingga berkontribusi dalam pengembangan sistem diagnosis berbasis *deep learning* yang lebih transparan dan dapat diandalkan.

2. METODE

Metode yang diusulkan dalam penelitian ini, mencakup 2 tahapan besar. Pada tahapan pertama dilakukan klasifikasi *X-ray* Dada untuk mendeteksi penyakit dengan kategori COVID-19, paru-paru normal, dan Pneumonia. Pada tahap kedua yaitu mengimplementasikan *explainable artificial intelligence* (XAI) menggunakan metode l*ayer-wise relevance propagation* (LRP) untuk menginterpretasikan hasil kinerja model.

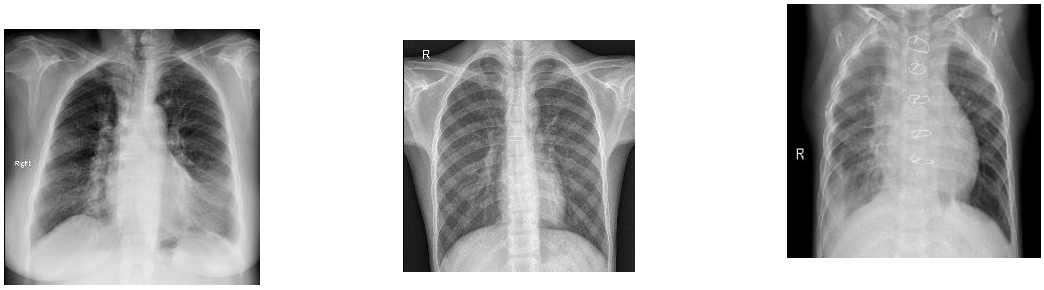


Gambar 1. Tahapan Penelitian

2.1 Dataset

Dataset yang digunakan dalam penelitian ini terdiri dari citra *X-ray* dada yang terbagi dalam tiga kategori, yaitu COVID-19, normal, dan Pneumonia. Seluruh citra diperoleh dari situs Mendeley Data yang menyediakan akses secara terbuka terhadap dataset ilmiah. Untuk tahap pelatihan dan evaluasi model *deep learning*, digunakan sebanyak 1000 citra untuk setiap kategori, sehingga total terdapat 3000 citra *X-ray*. Citra-citra tersebut disusun dalam folder terpisah berdasarkan kategorinya guna memudahkan pelabelan data.

Selanjutnya, pada tahap implementasi teknik interpretabilitas menggunakan *layer-wise relevance propagation* (LRP), digunakan 1 citra dari masing-masing kategori yang diambil dari sumber dataset yang sama, namun tidak disertakan dalam proses pelatihan maupun validasi sebelumnya. Hal ini dilakukan untuk menguji kemampuan model dalam memberikan interpretasi terhadap prediksi pada data baru yang belum pernah dilihat sebelumnya. Sehingga total keseluruhan data yang digunakan adalah 3003 citra.



(a) (b) (c)

Gambar 2. Citra *X-ray* (a) COVID-19 (b) Normal (c) Pneumonia.

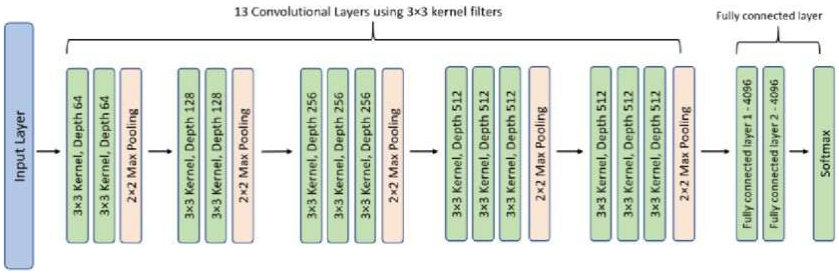
2.2 Preprocessing

Dalam penelitian ini, setiap citra *X-ray* dada terlebih dahulu melalui tahap *preprocessing* sebelum digunakan dalam proses pelatihan model. Tahap ini meliputi perubahan ukuran citra menjadi 224×224 piksel agar sesuai dengan format *input* arsitektur VGG16, konversi citra ke dalam format tensor, serta normalisasi berdasarkan nilai rata-rata (*mean*) dan standar deviasi (*standard deviation*) dari dataset ImageNet. Langkah-langkah ini bertujuan untuk memastikan bahwa data memiliki format dan skala yang konsisten dengan model pralatih, sehingga dapat mempercepat konvergensi dan meningkatkan stabilitas pelatihan [15].

2.3 Proses *Deep learning*

2.3.1 Visual Geometric Group 16

VGG16 merupakan arsitektur CNN yang teridiri terdiri atas 16 lapisan yang memiliki bobot yang dapat dilatih, yaitu 13 lapisan konvolusi dan 3 lapisan fully connected. Semua layer konvolusi dalam VGG16 menggunakan kernel berukuran kecil 3x3 dengan padding dan stride 1. Selain itu, di setiap beberapa lapisan konvolusi, digunakan operasi *max pooling* dengan ukuran 2x2 dan stride 2 untuk mengurangi dimensi spasial dari fitur. Serta dua lapisan *fully connected* masing-masing dengan 4096 unit, sebelum akhirnya diteruskan ke lapisan *fully connected* terakhir yang berfungsi sebagai *classifier* [16]. *Output* akhir dihasilkan oleh fungsi *softmax*, yang mengubah nilai-nilai akhir menjadi probabilitas dari masing-masing kelas target.



Gambar 3. Model arsitektur VGG16

2.3.2 Konfigurasi Pelatihan Model

Penelitian ini menggunakan arsitektur VGG16 dengan bobot pralatih dari ImageNet, di mana lapisan *feature extractor* dibekukan untuk mempertahankan kemampuan ekstraksi fitur umum, dan lapisan klasifikasi akhir diubah untuk menyesuaikan dengan jumlah kelas target. Strategi ini mengikuti pendekatan *transfer learning* yang telah terbukti efektif dalam klasifikasi citra medis karena dapat memanfaatkan pengetahuan dari domain sumber seperti ImageNet dan mengurangi waktu pelatihan serta kebutuhan data besar [17]. Untuk meningkatkan keragaman data pelatihan dan mencegah *overfitting*, dilakukan augmentasi data menggunakan *flip horizontal* dan *random rotation* [18]. Augmentasi ini terbukti efektif dalam meningkatkan akurasi klasifikasi citra *X-ray* sebagaimana ditunjukkan oleh studi di [19], yang menunjukkan bahwa *random rotation* dan *flipping horizontal* dapat membantu model belajar fitur yang lebih beragam. Model dilatih menggunakan *CrossEntropyLoss*, yang merupakan fungsi loss standar untuk klasifikasi multikelas, serta *Adam optimizer* yang dikenal efisien dalam mempercepat proses konvergensi dan adaptif terhadap skala gradien [20].

Pembekuan parameter konvolusional dan pelatihan ulang hanya pada lapisan akhir memungkinkan efisiensi pelatihan tanpa kehilangan performa dalam domain target. Penelitian ini, model dilatih menggunakan ukuran *batch* sebesar 32 dan jumlah *epoch* ditetapkan sebanyak 30. Pemilihan *batch size* 32 didasarkan pada pertimbangan keseimbangan antara stabilitas pelatihan dan efisiensi komputasi. Ukuran *batch* yang lebih kecil cenderung menghasilkan pembaruan gradien yang lebih sering dan dapat membantu model mencapai konvergensi yang lebih stabil, sementara ukuran *batch* yang terlalu besar dapat mengurangi kemampuan generalisasi model. Jumlah *epoch* sebanyak 30 dipilih untuk memberikan waktu pelatihan yang cukup agar model dapat mempelajari pola dari data tanpa mengalami *overfitting*. Untuk mengatasi *overfitting*, hanya model

terbaik yang disimpan, artinya selama fase pelatihan, jika akurasi validasi *epoch* lebih tinggi daripada akurasi tertinggi, maka model tersebut disimpan [21].

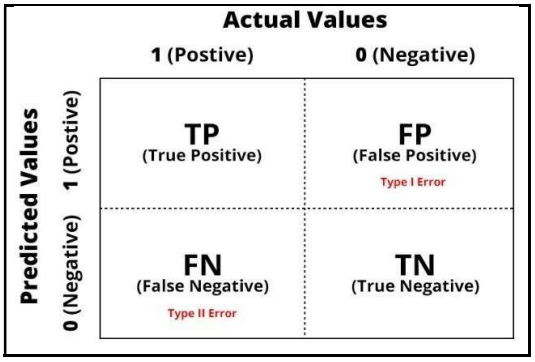
Eksplorasi *hyperparameter* dilakukan dengan menggunakan beberapa variasi *learning rate* dan pembagian dataset seperti yang disajikan pada tabel parameter pelatihan model.

Tabel 1. Parameter

Parameter	Nilai
Ratio train & val data	90:10, 80:20, 70:30
Optimizer	Adam
Activation Function	ReLU
Epoch	30
Batch size	32
Learning rate	0.01, 0.001, 0.0001, 0.00001

2.3.3 Evaluasi Model

Cara umum untuk mengevaluasi kinerja pengklasifikasian model adalah dengan menggunakan *confusion matrix*. *Confusion matrix* bekerja dengan merepresentasikan visual yang mengatur hasil prediksi model ke dalam tabel, dengan menunjukkan empat kombinasi nilai prediksi dan aktual yang berbeda [22]. Tabel ini mencakup:



Gambar 4. *Confusion Matrix*

Adapun penjelasan masing-masing istilah adalah sebagai berikut: *True Positive* (TP) merupakan jumlah data dengan kelas positif yang berhasil diklasifikasikan dengan benar oleh sistem. *True Negative* (TN) menunjukkan jumlah data dengan kelas negatif yang juga diklasifikasikan dengan tepat. *False Negative* (FN) adalah jumlah data yang sebenarnya termasuk kelas positif, namun salah diklasifikasikan sebagai negatif oleh sistem. Sebaliknya, *False Positive* (FP) mengacu pada jumlah data yang seharusnya termasuk kelas negatif, tetapi diklasifikasikan secara keliru sebagai positif oleh sistem. Ada beberapa rumus persamaan yang menggunakan nilai diatas, sebagai berikut:

a) *Accuracy*

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

b) *F1-Score*

$$F1 = 2\frac{Precision \times Recall}{Precision + Recall} \quad (3)$$

c) *Recall*

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

d) *Precision*

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

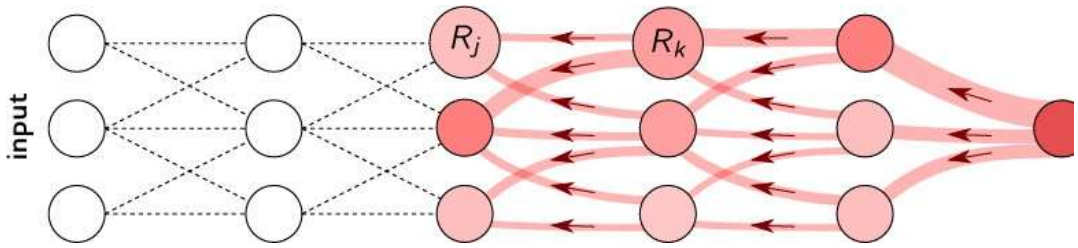
2.4 Implementasi XAI menggunakan LRP

2.4.1 Explainabel Artificial Intelligence

Explainable Artificial Intelligence (XAI) merujuk pada metodologi yang dirancang untuk membuat hasil dari model AI lebih transparan dan dapat dipahami oleh pengguna. Konsep ini menjadi sangat penting, terutama dalam aplikasi yang berkaitan dengan kesehatan, keuangan, dan pengambilan keputusan kritis lainnya, di mana pengguna perlu memercayai dan memahami bagaimana AI mencapai keputusan tertentu [23]. XAI bertujuan untuk mengkonversi model "black box", di mana keputusan tidak jelas, menjadi model "white box" yang lebih transparan, memungkinkan pengguna untuk memahami dan mengevaluasi hasil [24].

2.4.2 Layer-wise Relevance Propagation

Layer-wise Relevance Propagation (LRP) adalah teknik untuk interpretasi *output* model pembelajaran mendalam, khususnya untuk klasifikasi citra. Metode ini bekerja dengan menghitung kontribusi setiap *neuron* dalam jaringan saraf terhadap keputusan akhir dari model. Dengan menggunakan LRP, setiap piksel dari citra dapat dievaluasi untuk menentukan seberapa besar kontribusinya terhadap klasifikasi yang dihasilkan. Hal ini memberikan wawasan yang dalam mengenai proses pengambilan keputusan model dan memudahkan pengguna untuk memahami alasan di balik setiap prediksi yang dihasilkan oleh modebekerja dengan mendistribusikan ulang nilai relevansi yang terkait dengan *neuron* di lapisan atas ke lapisan *input* [25].



Gambar 5. Alur kerja LRP

Adapun cara kerja LRP sebagai berikut:

$$R_j = \sum_k \frac{z_{jk}}{\sum_j z_{jk}} \quad (5)$$

Teknik LRP bekerja dengan mendistribusikan ulang nilai relevansi yang terkait dengan *neuron* di lapisan *output* ke lapisan belakang *input*. Prosedur ini tunduk pada sifat konservasi, yang berarti bahwa relevansi yang diterima oleh lapisan tertentu harus didistribusikan ulang ke lapisan yang lebih rendah dalam jumlah yang sama. Penyebaran skor relevansi *Rk* pada lapisan tertentu ke *neuron* generik j dari lapisan yang lebih rendah dapat diperoleh dengan menerapkan aturan di mana *Rj* sesuai dengan skor relevansi pada *neuron j* karena *neuron k*. Kuantitas *zjk* memodelkan seberapa banyak *neuron j* telah berkontribusi untuk membuat *neuron k* relevan, sementara penyebut menegakkan properti konservasi. Cara penghitungan *zjk* dilambangkan sebagai aturan dan berbeda tergantung pada jenis lapisan yang terkandung dalam jaringan.

3. HASIL DAN PEMBAHASAN

Dalam penelitian ini, seluruh proses implementasi dilakukan menggunakan platform Google Colab dengan bahasa pemrograman Python serta pustaka *deep learning* PyTorch sebagai kerangka kerja utama. Pemilihan lingkungan ini didasarkan pada kemudahan akses, ketersediaan GPU untuk komputasi intensif, serta

kompatibilitas yang baik dengan berbagai pustaka pendukung yang diperlukan dalam pemodelan *deep learning* dan visualisasi interpretabilitas.

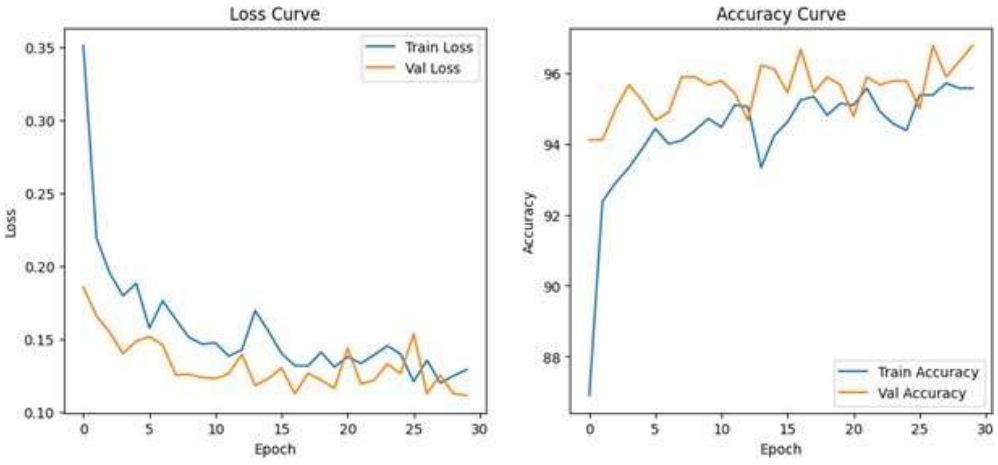
3.1 Hasil *Hyperparameter Testing*

Untuk memperoleh konfigurasi model terbaik, dilakukan serangkaian percobaan dengan memvariasikan rasio data latih dan validasi serta nilai learning rate. Tabel di bawah ini menyajikan hasil dari dua belas percobaan yang dilakukan, masing-masing menggunakan kombinasi rasio data (90:10, 80:20, dan 70:30) serta nilai learning rate (0.01, 0.001, 0.0001 dan 0.00001) dengan *batch size* 32 dan *epoch* 30.

Tabel 2. Hasil Pengujian

Percobaan ke-	Ratio Data	Learning rate	Accuracy	Val Accuracy	Loss	Val Loss
Percobaan ke-1	90:10	0,01	87.11%	94.67%	0.6525	0.1904
Percobaan ke-2	90:10	0,001	94.56%	96.33%	0.1551	0.0900
Percobaan ke-3	90:10	0,0001	94.52%	96.33%	0.1620	0.1304
Percobaan ke-4	90:10	0,00001	90.44%	92.33%	0.3122	0.2778
Percobaan ke-5	80:20	0,01	90.21%	96.00%	0.5045	0.2379
Percobaan ke-6	80:20	0,001	95.33%	96.50%	0.1400	0.1082
Percobaan ke-7	80:20	0,0001	95.04%	96.00%	0.1563	0.1394
Percobaan ke-8	80:20	0,00001	90.88%	93.33%	0.3218	0.2877
Percobaan ke-9	70:30	0,01	90.90%	95.44%	0.4815	0.2102
Percobaan ke-10	**70:30**	**0,001**	**95.57%**	**96.78%**	**0.1292**	**0.1117**
Percobaan ke-11	70:30	0,0001	90.00%	92.89%	0.3528	0.3113
Percobaan ke-12	70:30	0,00001	95.38%	96.50%	0.1400	0.1082

Berdasarkan hasil pengujian, diketahui bahwa model dengan rasio data 70:30 dan learning rate 0.001 (Percobaan ke-10) menghasilkan kinerja terbaik dengan akurasi validasi tertinggi sebesar 96.78% dan val loss terendah sebesar 0.1117.



Gambar 6. Grafik *loss* dan *accuracy* model terbaik

Visualisasi kurva loss pada pelatihan dan akurasi dengan kinerja terbaik pada pelatihan ke-10 menunjukkan proses pelatihan selama 30 *epoch* memberikan performa model yang stabil dan cukup optimal. Pada kurva loss, terlihat bahwa nilai *training loss* mengalami penurunan tajam dari awal pelatihan (~0.35) hingga mencapai nilai mendekati 0.11 di akhir *epoch*. Penurunan ini mengindikasikan bahwa model berhasil mempelajari pola dari data pelatihan secara bertahap. Sementara itu, *validation loss* menunjukkan nilai yang

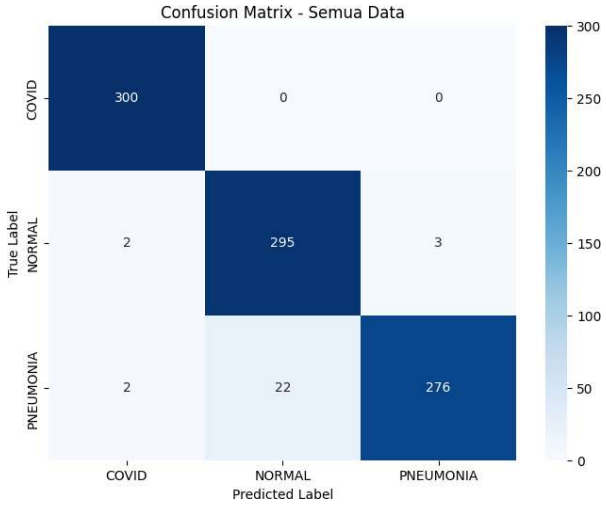
cenderung lebih rendah daripada *training loss* dan menurun secara stabil tanpa fluktuasi ekstrem, yang mengindikasikan tidak terjadinya *overfitting* secara signifikan.

Kurva akurasi pada pelatihan ke-10 memperlihatkan tren positif yang serupa. Akurasi pelatihan (*train accuracy*) meningkat secara bertahap dari sekitar 87% hingga mendekati 95%, sedangkan akurasi validasi (*val accuracy*) konsisten berada pada kisaran tinggi sejak awal pelatihan dan mencapai lebih dari 96% di akhir proses. Menariknya, akurasi validasi sempat lebih tinggi dibandingkan akurasi pelatihan, terutama pada awal dan pertengahan *epoch*. Hal ini dapat disebabkan oleh adanya teknik regularisasi, augmentasi data, atau distribusi data validasi yang lebih mudah diklasifikasikan. Secara keseluruhan, tidak terdapat perbedaan mencolok antara performa model pada data pelatihan dan validasi, sehingga dapat disimpulkan bahwa model memiliki kemampuan generalisasi yang baik dan tidak mengalami gejala *overfitting* maupun *underfitting* yang signifikan.

Berdasarkan kombinasi nilai *accuracy*, *val accuracy*, dan loss, maka percobaan ke-10 dengan rasio 70:30 dan learning rate 0.01 dipilih sebagai konfigurasi model terbaik untuk digunakan dalam proses pelatihan akhir dan analisis interpretabilitas dengan metode LRP.

3.2 Evaluasi Klasifikasi

Guna mengevaluasi performa model klasifikasi VGG16 dalam mendeteksi kondisi paru-paru berdasarkan citra X-ray, digunakan confusion matrix yang menggambarkan jumlah prediksi benar dan salah dari model terhadap masing-masing kelas: COVID-19, normal, dan Pneumonia. Confusion matrix berikut dihasilkan dari proses inferensi pada seluruh data uji:



Gambar 6. *Confusion Matrix*

Confusion matrix pada gambar 6 menunjukkan bahwa model mampu mengenali kelas COVID-19 dengan sangat baik yaitu dengan 300 data prediksi benar dari 300 data, sementara terdapat beberapa kesalahan dalam klasifikasi kelas normal dan Pneumonia**.** *Confusion matrix* dapat menghitung lebih detail kinerja pada tiap kelas menggunakan analisis metrik pada *precision*, *recall* dan *F1-score.*

Tabel 3. *Classification Report*

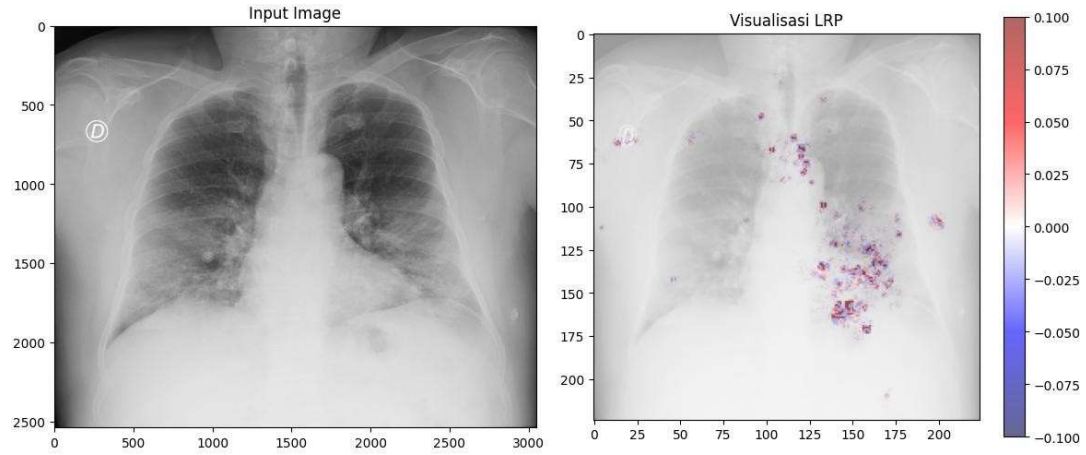
	Precision	Recall	F1-Score	Support
Covid	0.99	1.00	0.99	300
Normal	0.93	0.98	0.96	300
Pnemonia	0.99	0.92	0.95	300
Accuracy			0.97	900

Hasil evaluasi ditampilkan pada Tabel 3 menunjukkan bahwa model memiliki performa tertinggi pada kelas COVID-19 dengan nilai *precision* sebesar 0,99 dan *recall* sempurna 1,00, yang berarti model mampu mengenali seluruh citra COVID-19 tanpa kesalahan deteksi. Untuk kelas Normal, *precision* dan *recall* masing-masing adalah 0,93 dan 0,98, mengindikasikan performa tinggi namun masih terdapat beberapa prediksi salah. Kelas Pneumonia memperoleh *precision* sebesar 0,99 dan *recall* sebesar 0,92, menunjukkan bahwa meskipun prediksi positif sangat akurat, beberapa kasus Pneumonia masih terlewatkan oleh model. Secara keseluruhan, model mencapai akurasi sebesar 97%, dengan rata-rata *F1-score* sebesar 0,97. Hal ini menunjukkan bahwa model memiliki kinerja klasifikasi yang sangat baik secara keseluruhan dalam membedakan ketiga kategori penyakit.

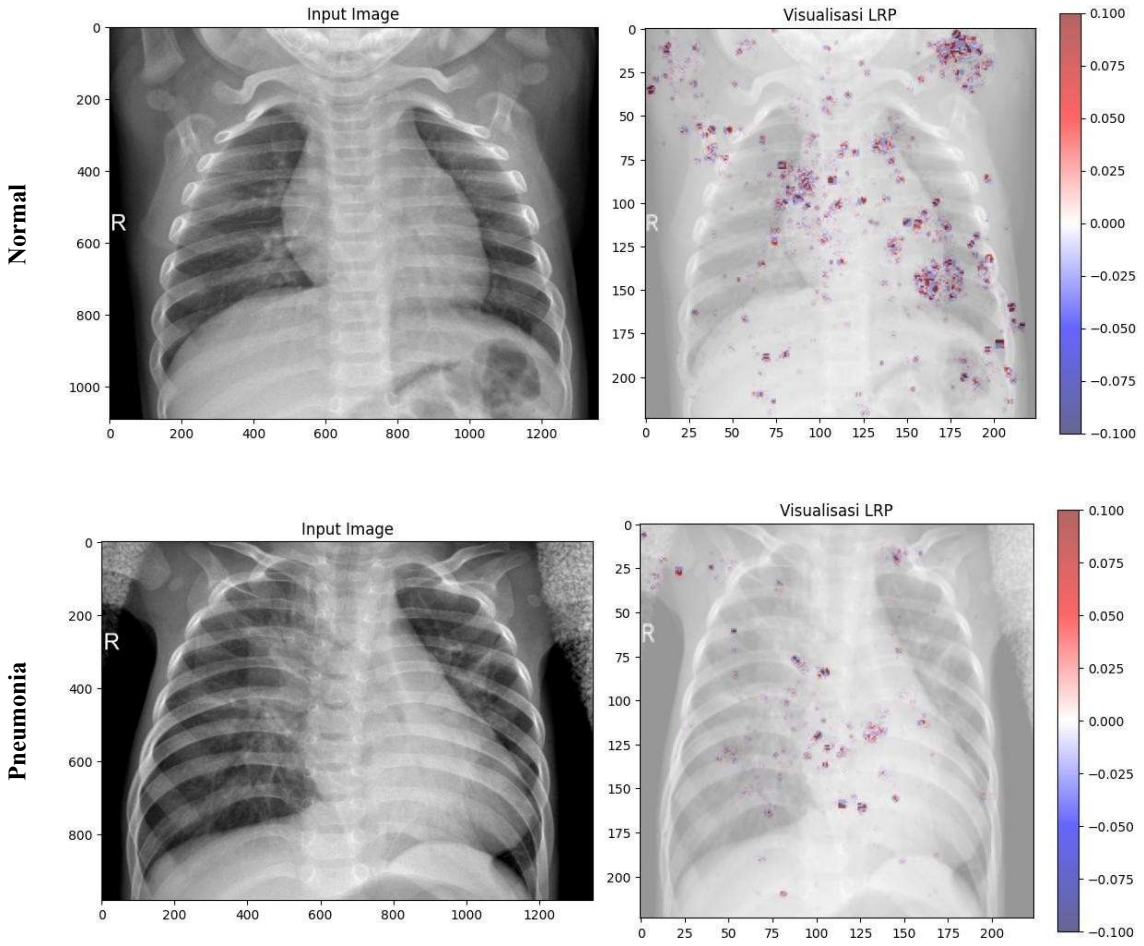
3.3 Implementasi LRP

Setelah model *deep learning* VGG16 berhasil dilatih untuk melakukan klasifikasi citra *X-ray* ke dalam tiga kategori COVID-19, normal, dan Pneumonia. Tahapan selanjutnya adalah melakukan interpretasi terhadap keputusan model menggunakan metode *layer-wise relevance propagation* (LRP). Implementasi LRP dilakukan untuk memperoleh gambaran visual mengenai bagian-bagian citra yang dianggap relevan oleh model dalam menghasilkan prediksi. Dalam proses ini, input dilakukan preprocessing untuk menyesuaikan model prediksi. Model yang telah dilatih kemudian dimuat ulang dengan parameter terbaik yang diperoleh selama pelatihan, selanjutnya dilakukan propagasi balik dari *output* hingga ke *input* layer untuk menghitung kontribusi setiap piksel terhadap prediksi yang dihasilkan.

Tabel 4. Hasil Implementasi LRP

	Visualisasi *input* dan *output*
COVID-19	

Visualisasi *input* dan *output*



Tabel 4 menampilkan hasil implementasi LRP pada model *deep learning* untuk klasifikasi citra *X-ray* dada. Visualisasi ini terdiri dari dua bagian yaitu gambar sebelah kiri (*input* gambar) adalah citra *X-ray* asli yang menjadi masukan ke model VGG16, sedangkan gambar disebelah kanan (visualisasi LRP) adalah hasil dari LRP yang menunjukkan area pada citra yang dianggap relevan dengan model dalam membuat keputusan. Area berwarna merah warna merah tua hingga merah muda (0.1 hingga 0.0) menunjukkan kontribusi positif mendukung keputusan yang diambil model. Sedangkan area berwarna biru tua hingga biru muda (0.0 hingga -0.1) menunjukkan kontribusi negatif muncul dari fitur yang membantah keputusaan model (kontribusi ke kelas lain). Area abu-abu atau tidak berwarna menunjukkan piksel yang tidak mempengaruhi prediksi model. Rentang nilai pada legenda adalah dari -0.1 hingga +0.1, yang merupakan rentang *relevance score* yang dihitung dari propagasi LRP.

Ditampilkan area berwarna biru dan merah seringkali berdampingan sehingga memberikan pemahaman yang membingungkan. Hal tersebut terjadi karena *layer convolution* dan *pooling* dalam CNN dapat menyebabkan piksel tetangga memiliki pengaruh yang berbeda terhadap fitur yang aktif. Sehingga, kontribusi berdekatan dari warna merah dan biru mengindikasikan area yang aktif secara signifikan, baik mendukung maupun menentang keputusan model. Selanjutnya artefak non-medis seperti huruf identifikasi D atau R yang terletak pada tepi citra turut memperoleh skor relevansi positif. Keikutsertaan artefak dalam peta relevansi mengindikasikan bahwa model mempelajari *shortcut features*, yaitu pola visual yang secara statistik berkorelasi dengan label namun tidak berkaitan langsung dengan patologi target.

Selain itu, pada citra kelas normal, terlihat bahwa area bahu kiri memperoleh skor relevansi positif yang cukup tinggi. Hal ini menunjukkan bahwa model menganggap area tersebut sebagai bagian yang relevan dalam pengambilan keputusan terhadap label normal. Pola ini kemungkinan muncul karena pada data pelatihan, area bahu kiri secara konsisten tampil bersih dari kelainan dan memiliki pencahayaan yang seragam, sehingga dipelajari sebagai ciri visual non-patologis. Dengan kata lain, model tidak hanya mengenali keberadaan fitur patologis, tetapi juga memanfaatkan ketiadaan kelainan sebagai indikator pendukung. Dalam konteks LRP, kontribusi positif dari area yang bersih ini memperkuat interpretasi bahwa model belajar berdasarkan statistik distribusi data, bukan semata pada makna medis. Temuan ini sekaligus menggarisbawahi pentingnya evaluasi visualisasi LRP secara menyeluruh untuk mengidentifikasi potensi bias atau shortcut learning dalam model.

4. KESIMPULAN

Berdasarkan hasil pembahasan, model klasifikasi yang dikembangkan menunjukkan akurasi validasi yang tinggi, yaitu mencapai 96.78%. Namun demikian, capaian akurasi yang tinggi tidak menjamin bahwa prediksi yang dihasilkan selalu benar secara semantik maupun klinis. Hasil interpretasi menggunakan metode *layer-wise relevance propagation* (LRP) mengungkapkan bahwa adanya sejumlah kasus di mana model memberikan prediksi yang benar dari segi label, namun tidak berdasarkan fitur-fitur relevan yang secara logis mendukung keputusan tersebut. Fenomena ini menunjukkan bahwa model berpotensi melakukan *shortcut learning*, yakni kecenderungan untuk mengandalkan pola atau artefak visual pada citra yang tidak memiliki relevansi medis, tetapi kebetulan sering muncul pada label tertentu.

Untuk mengatasi masalah tersebut, disarankan agar pelatihan model klasifikasi pada citra medis ke depannya menggunakan dataset telah dibersihkan dari artefak non-patologis, seperti tulisan, penanda, atau peralatan medis dalam gambar. Dengan cara ini, model dapat difokuskan untuk mempelajari karakteristik patologis yang signifikan, sehingga menghasilkan keputusan yang lebih adil, akurat, dan dapat diandalkan dalam konteks klinis yang sesungguhnya. Lebih lanjut, guna meningkatkan validitas klinis dari hasil interpretasi model, sangat disarankan agar visualisasi relevansi yang dihasilkan oleh LRP turut dievaluasi oleh tim medis, khususnya dokter spesialis radiologi atau pulmonologi. Evaluasi ini bertujuan untuk menilai kesesuaian antara area yang disorot oleh model dan area patologis yang secara klinis dianggap signifikan. Dengan melibatkan tinjauan ahli medis, interpretasi model tidak hanya akan lebih akurat, tetapi juga lebih dapat dipertanggungjawabkan dalam praktik klinis yang sesungguhnya.

REFERENSI

[1] D. Septhya *et al*., "Implementasi Algoritma Decision Tree Dan Support Vector Machine Untuk Klasifikasi Penyakit Kanker Paru," *Malcom Indonesian Journal of Machine Learning and Computer Science*, vol. 3, no. 1, pp. 15–19, 2023, doi: 10.57152/malcom.v3i1.591.

[2] N. Jain, A. Choudhury, J. Sharma, V. Kumar, D. De, and R. Tiwari, "A review of novel coronavirus infection (Coronavirus Disease-19)," *Global Journal of Transfusion Medicine*, vol. 5, no. 1, pp. 22–26, 2020, doi: 10.4103/GJTM.GJTM_24_20.

[3] M. drg. Oscar Primadi and M. dr. Anas Ma'ruf, "PROFIL KESEHATAN INDONESIA 2020," 2020.

[4] F. N. Azizah and D. Juniati, "Analisis Jenis Penyakit Paru-Paru Berdasarkan Chest X-Ray Menggunakan Metode Fuzzy C-Means," *Mathunesa Jurnal Ilmiah Matematika*, vol. 9, no. 2, pp. 322–331, 2021, doi: 10.26740/mathunesa.v9n2.p322-331.

[5] S. Saturi and S. Banda, "Advanced Lung Disease Detection and Classification Using Ge-U-Net-ODLwith Gabor Filters and Entropy-Based Feature Selection," *Journal of Sensors, IoT & Health Sciences*, vol. 2, no. 2, pp. 69–86, Jun. 2024, doi: 10.69996/jsihs.2024011.

[6] M. Kranthi, S. Sailaja, and E. Jyothi, "Deep Learning Approaches for Medical Image Processing in the Big Data Era," vol. 01, no. 01, pp. 24–31, 2024, doi: 10.58599/ijsmcse.2024.1108.

[7] M. E. Karar, E. E. Hemdan, and M. A. Shouman, "Cascaded Deep Learning Classifiers for Computer-Aided Diagnosis of COVID-19 and Pneumonia Diseases in X-Ray Scans," *Complex & Intelligent Systems*, vol. 7, no. 1, pp. 235–247, 2020, doi: 10.1007/s40747-020-00199-4.

[8] C. Karima, N. Bourbia, K. Messaoudi, and E.-B. Bourennane, "Analysis and Classification of Medical Images Using Deep Learning Algorithms," *Recent Advances in Computer Science and Communications*, vol. 18, Oct. 2024, doi: 10.2174/0126662558327739240925073925.

[9] M. M. Rahman, K. Matsuo, S. Matsuzaki, and S. Purushotham, "DeepPseudo: Pseudo Value Based Deep Learning Models for Competing Risk Analysis," *Proceedings of the Aaai Conference on Artificial Intelligence*, vol. 35, no. 1, pp. 479–487, 2021, doi: 10.1609/aaai.v35i1.16125.

[10] P. Gort, C. Claessens, F. van der Sommen, and P. H. N. de With, "Evaluating the Interpretability of Prototype Networks for Medical Image Analysis," p. 54, 2025, doi: 10.1117/12.3046678.

[11] D. A. Safitri, Y. P. M. Mamesah, and J. F. J. Timban, "Gambaran Foto Toraks Pada Pasien Tuberkulosis Paru Dengan Penyakit Ginjal Kronik Di RSUP Prof. Dr. R. D. Kandou Periode Januari-Juni 2022," *Medical Scope Journal*, vol. 4, no. 1, pp. 93–98, 2023, doi: 10.35790/msj.v4i1.44722.

[12] D. Kim, J. Lee, J. Moon, and T. Moon, "Interpretable Deep Learning-based Hippocampal Sclerosis Classification," *Epilepsia Open*, vol. 7, no. 4, pp. 747–757, 2022, doi: 10.1002/epi4.12655.

[13] S. Gupta *et al.*, "Four Transformer-Based Deep Learning Classifiers Embedded with an Attention U-Net-Based Lung Segmenter and Layer-Wise Relevance Propagation-Based Heatmaps for COVID-19 X-ray Scans," *Diagnostics*, vol. 14, no. 14, p. 1534, Jul. 2024, doi: 10.3390/diagnostics14141534.

[14] F. Ahmed *et al.*, "Identification of kidney stones in KUB X-ray images using VGG16 empowered with explainable artificial intelligence," *Sci Rep*, vol. 14, no. 1, p. 6173, Mar. 2024, doi: 10.1038/s41598-024-56478-4.

[15] L. Sağır, E. Kaba, M. H. Yiğit, F. Taşçı, and H. Uzun, "Predicting Semen Analysis Parameters From Testicular Ultrasonography Images Using Deep Learning Algorithms: An Innovative Approach to Male Infertility Diagnosis," *J Clin Med*, vol. 14, no. 2, p. 516, 2025, doi: 10.3390/jcm14020516.

[16] D. Shen, G. Wu, and H.-I. Suk, "Deep Learning in Medical Image Analysis.," *Annu Rev Biomed Eng*, vol. 19, pp. 221–248, Jun. 2017, doi: 10.1146/annurev-bioeng-071516-044442.

[17] H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt, "Transfer learning for medical image classification: a literature review," *BMC Med Imaging*, vol. 22, no. 1, p. 69, Dec. 2022, doi: 10.1186/s12880-022-00793-7.

[18] R. Hao, K. Namdar, L. Liu, M. A. Haider, and F. Khalvati, "A Comprehensive Study of Data Augmentation Strategies for Prostate Cancer Detection in Diffusion-Weighted MRI Using Convolutional Neural Networks," *J Digit Imaging*, vol. 34, no. 4, pp. 862–876, Aug. 2021, doi: 10.1007/s10278-021-00478-7.

[19] B. Liu, C. Tan, S. Li, J. He, and W. Hong-yan, "A Data Augmentation Method Based on Generative Adversarial Networks for Grape Leaf Disease Identification," *Ieee Access*, vol. 8, pp. 102188–102198, 2020, doi: 10.1109/access.2020.2998839.

[20] H. Iiduka, "Appropriate Learning Rates of Adaptive Learning Rate Optimization Algorithms for Training Deep Neural Networks," 2020, doi: 10.48550/arxiv.2002.09647.

[21] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jul. 2017, pp. 2261–2269. doi: 10.1109/CVPR.2017.243.

[22] S.-C. Kim and Y. Cho, "Predictive System Implementation to Improve the Accuracy of Urine Self-Diagnosis With Smartphones: Application of a Confusion Matrix-Based Learning Model Through RGB Semiquantitative Analysis," *Sensors*, vol. 22, no. 14, p. 5445, 2022, doi: 10.3390/s22145445.

[23] A. K. M. Nor, S. R. Pedapati, M. Muhammad, and V. Leiva, "Overview of Explainable Artificial Intelligence for Prognostic and Health Management of Industrial Assets Based on Preferred Reporting Items for Systematic Reviews and Meta-Analyses," *Sensors*, vol. 21, no. 23, p. 8020, 2021, doi: 10.3390/s21238020.

[24] S. Lebovitz, H. Lifshitz-Assaf, and N. Levina, "To Engage or Not to Engage With AI for Critical Judgments: How Professionals Deal With Opacity When Using AI for Medical Diagnosis," *Organization Science*, vol. 33, no. 1, pp. 126–148, 2022, doi: 10.1287/orsc.2021.1549.

[25] P. R. A. S. Bassi, S. S. J. Dertkigil, and A. Cavalli, "Improving deep neural network generalization and robustness to background bias via layer-wise relevance propagation optimization," *Nat Commun*, vol. 15, no. 1, p. 291, 2024, doi: 10.1038/s41467-023-44371-z.