SCIENCE – FUTURE of LITHUANIA

Information technologies and multimedia
Informacinės technologijos ir multimedija

DETECTION OF PNEUMONIA FROM CHEST X-RAYS USING CONVOLUTIONAL NEURAL NETWORKS

Paulius BUNDZA [iD], Justas TRINKŪNAS [iD][✉]

Vilnius Gediminas Technical University, Vilnius, Lithuania

Abstract. Pneumonia detection from chest X-rays is crucial for early diagnosis, and deep learning models – specifically convolutional neural networks (CNNs) – have shown promise in automating this process. In this study, a CNN using the DenseNet-121 architecture was developed and trained, referred to as LDCS2, to classify chest X-ray images as pneumonia or normal, using a combined dataset from three publicly available sources. The CNN approach was chosen over Vision Transformers (ViT) due to lower computational requirements and better performance with limited data. A traditional training, validation, and testing split was used instead of k-fold cross-validation to reduce execution time. LDCS2 demonstrated excellent discrimination between pneumonia and normal images alongside high computational efficiency. These findings highlight the potential of DenseNet-based CNNs for automated pneumonia diagnosis, particularly in resource-constrained settings.

1. Introduction

Pneumonia is a severe respiratory infection characterized by inflammation of the lung tissues, commonly caused by bacteria, viruses, or fungi. It remains a leading cause of morbidity and mortality worldwide, especially among vulnerable populations such as children under five years old, elderly individuals, and immunocompromised patients (World Health Organization, 2023). Early diagnosis of pneumonia is critical, as delayed treatment can lead to severe complications or death (Katz & Williams, 2021). Chest radiography (X-ray) remains the most accessible and frequently utilized diagnostic imaging technique due to its relative affordability, rapid acquisition, and broad availability across various healthcare settings (Rajpurkar et al., 2017). Nevertheless, interpreting chest X-rays for pneumonia remains challenging due to variations in disease presentation, overlapping anatomical structures, and subtle pathological signs that can be difficult to discern even for experienced radiologists (Rajpurkar et al., 2018; Sait et al., 2022). These challenges result in considerable diagnostic inconsistency and inter-observer variability, emphasizing the need for reliable and automated diagnostic solutions (Sait et al., 2022).

Recent advancements in deep learning, particularly convolutional neural networks (CNNs), have shown exceptional potential in medical image analysis, demonstrating accuracy comparable to or surpassing human experts in various diagnostic tasks (Irvin et al., 2019; Rajpurkar et al., 2018). CNNs excel at extracting relevant visual features from imaging data, significantly improving the accuracy and consistency of pneumonia diagnosis from chest X-rays (Irvin et al., 2019). Although Vision Transformers (ViT) have recently emerged as powerful image classification tools, their effectiveness heavily depends on large datasets and substantial computational resources, making CNNs more practical for medical applications where resources may be limited (Dosovitskiy et al., 2021).

In this study, DenseNet-121 CNN architecture was utilized due to its proven capability and efficiency in medical image classification tasks, including chest X-ray analysis (Irvin et al., 2019; Rajpurkar et al., 2018). To enhance the robustness and generalizability of the model, three publicly available datasets were integrated: the Kaggle Chest X-Ray Pneumonia dataset (Mooney, 2018), the COVID-19 Radiography Database (Rahman et al., 2021b), and the Curated Dataset for COVID-19 Posterior-Anterior Chest Radiography Images (X-rays) dataset (Sait et al., 2022).

Combining these diverse datasets aims to provide comprehensive coverage of various clinical manifestations of pneumonia.

Traditional data splitting strategy was choosen (training, validation, and testing subsets) instead of k-fold cross-validation to reduce computational requirements and execution time, aligning with practical constraints commonly encountered during model development (Vabalas et al., 2019). The goal of this research is to assess the effectiveness of CNN-based approaches for pneumonia detection from chest X-rays, focusing on their applicability in real-world clinical scenarios, particularly in resource-limited healthcare settings.

2. Related work

2.1. CNNs for pneumonia detection

CNNs have revolutionized pneumonia detection from chest radiographs due to their exceptional ability to learn discriminative features directly from images. Several CNN architectures, notably ResNet, DenseNet, and EfficientNet, have been extensively explored. ResNet architectures have shown remarkable performance, achieving accuracies exceeding 90% and AUC-ROC scores around 0.97 (Rajpurkar et al., 2017). DenseNet, particularly DenseNet-121, has gained attention due to its efficient use of parameters and high diagnostic performance, often reporting sensitivity and specificity above 90% (Irvin et al., 2019; Rajpurkar et al., 2018). EfficientNet models further advanced CNN capabilities by optimizing network depth, width, and resolution, resulting in even higher accuracy and robustness in pneumonia detection tasks (Sahiner et al., 2019; Tan & Le, 2019).

2.2. Vision transformers (ViTs) in medical imaging

Vision Transformers (ViTs), a more recent development in medical imaging, have introduced an alternative approach by leveraging self-attention mechanisms. ViTs have been successfully applied to chest X-ray classification, with studies reporting comparable or superior performance to CNNs in terms of accuracy, sensitivity, and specificity. For example, Shamshad et al. (2023) demonstrated ViT models achieving accuracies up to 97.6%, with sensitivity of 95.4% and specificity of 98.1% in pneumonia detection tasks. Similarly, Dosovitskiy et al. (2021) showed that ViTs could achieve performance comparable to the best CNN architectures, reporting accuracy rates around 96%. However, their higher computational demands, including increased training time and substantial GPU memory requirements, remain significant drawbacks. These computational and data-intensive requirements limit their practical deployment, particularly in resource-constrained healthcare environments (Han et al., 2022; Shamshad et al., 2023).

2.3. Traditional machine learning approaches

Traditional machine learning approaches, such as Support Vector Machines (SVMs), Random Forests, and k-Nearest Neighbors (k-NN), historically played a prominent role in pneumonia detection tasks. These methods typically rely on manually engineered features such as texture descriptors, shape features, and radiomic parameters extracted from chest X-rays. Varshni et al. (2019) compared traditional classifiers like SVMs and Random Forests with deep CNN methods, reporting that traditional approaches often achieved lower accuracy, generally ranging from 70% to 85%. In a comparative study, SVM classifiers using hand-crafted features achieved accuracy rates around 85–90%, significantly lower than deep CNN methods (Elshazly et al., 2020; Varshni et al., 2019). Furthermore, SVMs coupled with CNN-derived deep features slightly improved performance, achieving accuracies around 91%, but still did not surpass fully CNN-based approaches. Thus, deep learning methods, particularly CNNs, have become preferred due to their superior automatic feature extraction capabilities, consistently higher sensitivity, specificity, and overall diagnostic accuracy (Varshni et al., 2019).

2.4. Benchmark datasets and evaluation metrics

Benchmark datasets have played a crucial role in evaluating pneumonia detection models by providing standardized platforms for performance comparisons. Prominent datasets include the Kaggle Chest X-ray Pneumonia dataset (Mooney, 2018), consisting of approximately 5.863 pediatric chest radiographs labeled as pneumonia or normal; the COVID-19 Radiography Database (Rahman et al., 2021b), comprising 42.335 radiographs labeled as COVID-19, other pneumonia types, and normal cases; and the Curated Dataset for COVID-19 Posterior-Anterior Chest Radiography Images (X-rays) dataset (Sait et al., 2022), which includes 9.209 pediatric chest radiographs labeled for pneumonia research. These datasets have significantly contributed to consistent performance comparisons. For instance, models evaluated on the Kaggle Chest X-ray Pneumonia dataset frequently report accuracy rates ranging from 95% to 98%, sensitivity rates exceeding 95%, and AUC-ROC scores typically approaching 0.99 (Irvin et al., 2019; Sahiner et al., 2019). Additionally, models validated on the COVID-19 Radiography Database have achieved accuracy around 97–99%, demonstrating robust performance in multi-class scenarios involving COVID-19 pneumonia differentiation (Rahman et al., 2021b).

2.5. Real-world deployment and challenges

Despite impressive performance in controlled experiments, real-world deployment of AI-based pneumonia detection systems faces several challenges. Data scarcity, variability across different patient populations, and differences in imaging protocols significantly affect model generalizability

(2020; Roberts et al., 2021). Additionally, issues related to interpretability and explainability remain barriers to clinical adoption, as healthcare providers require transparency in model predictions for decision-making (Rudin, 2019; Shamshad et al., 2023). Moreover, computational limitations and resource constraints in clinical environments further complicate widespread implementation (Han et al., 2022). Addressing these challenges through robust model validation, explainable AI techniques, and resource-aware optimization is essential for successful integration into clinical workflows (Roberts et al., 2021; Shamshad et al., 2023).

2.6. LDCS2 comparison with existing approaches

Comparing the LDCS2 model developed in this research to existing models described in recent literature demonstrates notable performance enhancements. Moujahid et al. (2020) utilized pre-trained CNN models, achieving a maximum accuracy of 96.81%, sensitivity of 96.5%, and specificity of 96.8% using the VGG16 model, which are slightly lower than the metrics achieved by LDCS2.

Similarly, studies by Hassantabar et al. (2020) and Das et al. (2021), focusing on binary classification for COVID-19 detection, reported accuracies of 93.2% and 95.58%,

respectively, both of which are significantly lower compared to LDCS2 model's accuracy of 97.13%. The precision and recall scores reported by Das et al. (2021) were also lower, further highlighting the superior performance of LDCS2 model based on DenseNet-121 architecture.

The multi-class classification studies conducted by Ibrahim et al. (2024) and Wang and Zhang (2022) achieved high accuracies of 98.05% and 83.3%, respectively. Although Ibrahim et al. (2024) reported higher accuracy, their dataset included four distinct disease classes, making direct comparison challenging due to the inherently increased complexity and lower baseline performance typically seen in multi-class classification tasks.

Moreover, Albahli et al. (2021) tested DenseNet-121 for multi-class lung disease detection with significantly lower accuracy (41.02%), emphasizing the complexity associated with multi-class scenarios. The results presented in this research thus demonstrate that the developed DenseNet-121 model provides superior performance specifically for binary pneumonia classification tasks, outperforming or closely matching state-of-the-art models referenced in recent literature (Abbas et al., 2021; Podder et al., 2021).

Below in Table 1, full comparison with existing approaches can be seen.

Table 1. LDCS2 comparison with existing approaches

Reference	Main research question/problem	Used approach	Dataset used	Attributes used for prediction	Main result
Podder et al. (2021)	Can Mask R-CNN effectively classify COVID-19 and Non-COVID-19 cases?	Mask R-CNN with region-based feature extraction	COVID-19 X-ray dataset (publicly available)	2 classes (COVID-19, Not-COVID-19)	Accuracy: 96.98%
Moujahid et al. (2020)	How effective are CNN models in distinguishing Normal and Pneumonia cases?	Pre-trained models (VGG16, VGG19, NasNetMobile, ResNet152V2, InceptionResNetV2)	Public chest X-ray dataset (e.g., Kaggle Chest X-ray Pneumonia dataset)	2 classes (Normal, Pneumonia)	VGG16 achieved Accuracy: 96.81%, Sensitivity: 96.5%, Specificity: 96.8%
Das et al. (2021)	Can CNN-based models classify Normal and COVID-19 cases with high accuracy?	Combined model using InceptionV3, ResNet50V2, DenseNet201	Public COVID-19 dataset (2,905 images)	2 classes (Normal, COVID-19)	ResNet50V2 achieved Accuracy: 95.58%, Precision: 95.91%, Recall: 95.11%
Irmak (2021)	How can CNN models classify COVID-19 severity levels (Mild, Moderate, Severe, Critical)?	Novel CNN architecture with custom layers for severity classification	COVID-19 X-ray dataset (publicly available)	4 classes (COVID-19 (Mild, Moderate, Severe, Critical))	Accuracy: 83.4%, demonstrating potential for severity classification
Hassantabar et al. (2020)	Can CNN models accurately classify Normal and COVID-19 cases?	Deep Neural Network (DNN) combined with CNN	Chest X-ray datasets (e.g., COVID-19 Radiography Database)	2 classes (Normal, COVID-19)	Accuracy: 93.2%, Sensitivity: 92.8%, Specificity: 93.5%
Ismael and Şengür (2021)	How effective is ResNet50+SVM in classifying Normal and COVID-19 cases?	Fine-tuned ResNet50 combined with Support Vector Machine (SVM)	Public COVID-19 dataset (e.g., COVIDx dataset)	2 classes (Normal, COVID-19)	Accuracy: 94.7%, Precision: 94.5%, Recall: 94.9%

Reference	Main research question/problem	Used approach	Dataset used	Attributes used for prediction	Main result
Wang and Zhang (2022)	Can MARnet classify Nodules, Atelectasis, Infection, and Normal cases effectively?	MARnet with multi-class classification layers	Five different datasets (X-rays and CT scans)	4 classes (Nodules, Atelectasis, Infection, Normal)	Accuracy: 83.3%, demonstrating uniformly high performance across all classes
Ibrahim et al. (2024)	How effective are CNN models in classifying Normal, COVID-19, Pneumonia, and Lung Cancer cases?	VGG19+CNN, ResNet152V2+GRU, ResNet152V2+Bi-GRU	Chest X-ray and CT scans (e.g., COVID-19 CT and X-ray datasets)	4 classes (Normal, COVID-19, Pneumonia, Lung cancer)	VGG19+CNN achieved Accuracy: 98.05%, Sensitivity: 97.8%, Specificity: 98.3%
Podder et al. (2021)	What is the performance of multi-class CNN models in lung disease classification?	Hybrid CNN models (Vanilla Gray, Vanilla RGB, VGG, VDSNet, Modified CapsNet, Basic CapsNet)	Chest X-ray datasets (e.g., NIH Chest X-ray dataset)	15 classes (No finding, Infiltration, Mass, Effusion, Atelectasis, Nodule, Pneumothorax, Consolidation, Plural thickening, Hernia, Cardiomegaly, Emphysema, Edema, Fibrosis, Pneumonia)	VDSNet achieved Accuracy: 73%, emphasizing recall for early detection of lung diseases
Rahman et al. (2021a)	Can CNN + HOG classify Normal, Pneumonia, and COVID-19 cases accurately?	CNN combined with Histogram of Oriented Gradients (HOG)	COVID-19 X-ray dataset (publicly available)	3 classes (Normal, Pneumonia, COVID-19)	Accuracy: 92.95%, Sensitivity: 92.5%, Specificity: 93.1%
Abbas et al. (2021)	How effective are DeTrac models in classifying Normal, COVID-19, and SARS cases?	DeTrac ensemble models (AlexNet, VGG19, ResNet, GoogleNet, SqueezeNet)	Chest X-ray datasets (e.g., COVID-19 Radiography Database)	3 classes (Normal, COVID-19, SARS)	DeTrac+VGG19 achieved Accuracy: 97.53%, Sensitivity: 97.2%, Specificity: 97.8%
Albahli et al. (2021)	What is the performance of DenseNet121 and InceptionResNetV2 in multi-class lung disease classification?	DenseNet121, InceptionResNetV2, ResNet152V2	Chest X-ray datasets (e.g., NIH Chest X-ray dataset)	14 classes (Atelectasis, Cardiomegaly, Consolidation, Edema, Mass, Effusion, Emphysema, Nodule, Pneumothorax, Fibrosis, Infiltration, Plural thickening, Pneumonia, Normal)	InceptionResNetV2 achieved Accuracy: 41.02%, demonstrating limited performance for complex multi-class tasks
Proposed approach	How effectively can a DenseNet-121-based CNN (the LDCS2 model) detect pneumonia in chest X-rays, while remaining computationally efficient for resource-limited environments?	DenseNet-121 (Normal vs. Pneumonia) + balanced data, extensive augmentation, and AdamW (AMSGrad) with LR scheduling & early stopping	Three public datasets combined, yielding 20,550 balanced images split into train/val/test	2 classes. Chest X-ray images resized to 224×224, focusing on binary classification (Normal vs. Pneumonia)	Test Accuracy: 97.13% Sensitivity: 98.69% Specificity: 95.58% Precision: 99.87% F1-score: 99.27% ROC-AUC: 0.9958

3. Methodology

3.1. Data collection for LDCS2

The study leveraged three publicly available chest X-ray datasets for pneumonia detection:

Kaggle Chest X-ray Pneumonia dataset, containing 5.863 pediatric radiographs labeled as normal or pneumonia (Mooney, 2018).

COVID-19 Radiography Database, consisting of 42.335 images classified as COVID-19, other pneumonia, or normal (Rahman et al., 2021b).

Curated Dataset for COVID-19 Posterior-Anterior Chest Radiography Images (X-rays) dataset, comprising 9.209 pediatric radiographs for pneumonia classification (Sait et al., 2022).

Below Figure 1 depicts normal and pneumonia affected lungs X-ray sample images from Kaggle Chest X-ray Pneumonia dataset (Mooney, 2018).
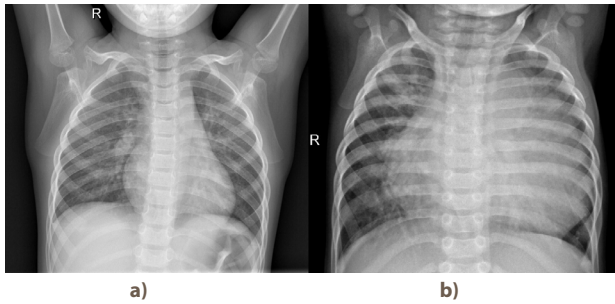
Figure 1. X-ray samples from Kaggle Chest X-ray Pneumonia dataset: a) normal lungs; b) pneumonia affected lungs (source: Mooney, 2018)

The combined dataset initially contained 57.407 labeled chest X-ray images. However, due to class imbalance across datasets, a balancing procedure was applied to ensure equal representation of pneumonia and normal cases. The final balanced dataset contained 20.550 images, equally distributed between pneumonia-positive and normal cases.

As shown in Table 2 the dataset was split into three subsets and then balanced:

Training Set: 67% of the images (28.140) used to train the model.

Validation Set: 20% of the images (8.400) used for hyperparameter tuning.

Test Set: 13% of the images (5.460) reserved for final model evaluation.

Table 2. LDCS2 data balancing

Split	Normal	Pneumonia	Total
Training	8.937	6.018	14.955
Validation	2.870	2.355	5.225
Test	3.239	1.902	5.141
Balanced training	6.018	6.018	12.036
Balanced validation	2.355	2.355	4.710
Balanced test	1.902	1.902	3.804

3.2. LDCS2 data preprocessing and augmentation

To effectively train the DenseNet-121 model for accurate pneumonia detection, all input images underwent standardized preprocessing to ensure uniformity across the dataset. Each image was resized to a fixed dimension of 224×224 pixels to align with the input requirements of the DenseNet-121 architecture. Additionally, images were normalized using parameters derived from the ImageNet dataset, specifically setting the pixel intensities to a mean of [0.485, 0.456, 0.406] and a standard deviation of [0.229, 0.224, 0.225].

To enhance the robustness and generalization capabilities of the model, extensive data augmentation techniques were employed during the training phase. The augmentations included:

Random affine transformations: Images underwent random scaling between 0.85 to 1.15, translations within ±10% of the original dimensions, and rotations within ±20°, thereby improving the model's invariance to orientation and positional variations.

Random brightness and contrast adjustments: Pixel intensities were randomly modified by up to ±20%, helping the model to become resilient to variations in image acquisition settings.

Horizontal flips: Implemented with a probability of 50%, these flips contributed to the diversity of the dataset, reducing directional bias.

Gaussian noise addition: Applied with a 30% probability to simulate real-world imaging conditions, such as sensor noise or low-quality captures.

Gaussian blur: Introduced with a 20% probability, this augmentation emulated scenarios of suboptimal imaging focus or movement artifacts.

Grid distortion: Executed with a probability of 30%, grid distortion introduced subtle spatial variations, aiding the model to better generalize to different anatomical alignments and orientations.

Mixup augmentation: Implemented with a 50% probability and an alpha parameter of 0.2, Mixup augmentation helped in enhancing model calibration and regularization by generating new training examples through linear combinations of existing image pairs.

Validation and testing images were subjected to minimal preprocessing, involving resizing to 224×224 and normalization only, thereby preserving their original characteristics to accurately evaluate the model's true predictive performance. Additionally, during inference, test-time augmentation (TTA) was conducted, involving five distinct transforms that included horizontal flips and minor rotations, ensuring predictions were robust against minor variations in input data.

This comprehensive preprocessing and augmentation strategy, as summarized in Table 3, contributed significantly to the model's effectiveness in distinguishing between normal and pneumonia-affected images, ultimately supporting higher validation and test accuracies.

Table 3. LDCS2 data augmentation techniques

Technique	Value
Random Affine	(scale: 0.85–1.15, translate: ±10%, rotate: ±20°)
Random Brightness/Contrast	(±20%)
Horizontal Flip	(50% probability)
Gaussian Noise	(30% probability)
Gaussian Blur	(20% probability)
Grid Distortion	(30% probability)
Mixup Augmentation	(50% probability, alpha = 0.2)
Test-Time Augmentation	5 transforms with flips and small rotations

3.3. LDCS2 model architecture

The implemented deep learning model utilizes DenseNet-121 as its backbone, selected specifically for its efficient handling of parameters and robust feature extraction performance. DenseNet-121 is advantageous due to its densely connected convolutional layers, which alleviate the vanishing gradient problem, enhance feature propagation, and significantly improve gradient flow through the network.

The chosen DenseNet-121 architecture comprises a total of 6,955,906 trainable parameters, ensuring the model's complexity remains manageable while still capturing intricate patterns present in medical imaging data effectively. This relatively modest parameter count promotes computational efficiency without compromising the accuracy and precision necessary for medical image classification tasks.

To further enhance model performance and reduce the risk of overfitting, a customized classification head was integrated into the base architecture. This classification head incorporates a dropout layer with a dropout rate of 0.2, strategically implemented to randomly deactivate neurons during training, thus increasing model generalizability. Following dropout, a fully connected linear layer was applied, transforming the extracted features into the final class predictions. Specifically, this linear layer maps 1024 input features directly to the two defined classes: NORMAL and PNEUMONIA. The total number of trainable parameters within this structure amounts to 6,955,906, all of which are fine-tuned during the training process.

In summary, the customized DenseNet-121 model employed in this study, as summarized in Table 4, effectively balances computational demands with classification accuracy, ensuring suitability for practical deployment in clinical environments.

Table 4. LDCS2 model architecture

Layer	Output shape	Parameters
Input Layer	(224, 224, 3)	0
DenseNet-121 Base	(7, 7, 1024)	6,952,832
Global Average Pooling	(1024)	0
Dropout (0.2)	(1024)	0
Dense (Output)	(2)	2,050

3.4. LDCS2 training strategy

The DenseNet-121 model training incorporated advanced optimization strategies aimed at maximizing performance and reducing the risk of overfitting. The AdamW optimizer was selected due to its effectiveness in handling large-scale neural networks and its inherent capability to manage weight decay directly, enhancing generalization. Specifically, the AMSGrad variant of AdamW was utilized (amsgrad = True) to stabilize convergence during training. An initial learning rate of 0.00020 was chosen to ensure a balance between fast conver-

gence and model stability, with weight decay configured at 0.00010 to further regularize the learning process and prevent overfitting.

Learning rate scheduling was dynamically adjusted using the ReduceLROnPlateau algorithm, which decreased the learning rate by a factor of 0.5 after 5 epochs without improvement in validation loss, facilitating finer convergence as the model approached optimal parameters.

The training batch size was set at 64, providing a balanced trade-off between memory consumption and gradient estimation stability. This size allowed optimal utilization of the GPU's memory resources, peaking at approximately 4.56 GB usage per epoch, ensuring efficient GPU utilization without excessive memory overhead.

To enhance the robustness of training and performance, a combination of data augmentations was applied alongside mixup augmentation ($\alpha = 0.2$) with a 50% probability. Early stopping was configured with a patience of 15 epochs, halting training if validation accuracy failed to improve, thus conserving computational resources and avoiding unnecessary epochs.

Overall, these meticulous optimization choices, as summarized in Table 5, were essential for ensuring efficient resource utilization and maximizing the predictive capabilities of the DenseNet-121 model for pneumonia detection.

Table 5. Proposed model hyperparameters

Parameter	Value	Description
Batch Size	64	Number of images processed in each training step
Initial Learning Rate	0.00020 (2e-4)	Starting learning rate
Optimizer	AdamW	Optimizer with weight decay and amsgrad = True
Weight Decay	0.00010	L2 regularization factor
LR Scheduler	ReduceLROnPlateau	Reduces learning rate by factor of 0.5 with 5 epochs patience
Dropout Rate	0.2	Dropout probability for regularization
Loss Function	CrossEntropyLoss	With class balancing
Early Stopping	15	Patience in epochs before stopping training
Mixup Alpha	0.2	Alpha parameter for mixup augmentation
Max Epochs	100	Maximum number of training epochs

3.5. LDCS2 evaluation metrics

The model's effectiveness in classifying pneumonia cases from chest X-ray images was comprehensively evaluated using multiple standard metrics designed to assess its diagnostic accuracy and clinical applicability. The primary metrics included:

Accuracy: This measures the proportion of total correct predictions (both true positives and true negatives) out of all predictions. The model demonstrated an impressive accuracy of 97.13% on the test dataset, indicating a high proportion of correct predictions across both classes.

Sensitivity (Recall): Sensitivity, a critical metric in medical diagnostics, represents the model's ability to correctly identify patients with pneumonia. The model achieved a high sensitivity of 98.69%, indicating robust performance in correctly diagnosing true positives and effectively minimizing false-negative cases.

Specificity: Specificity measures the model's ability to correctly recognize individuals who do not have pneumonia. Achieving 95.58% specificity demonstrates that the model effectively minimizes false-positive rates, providing confidence in negative diagnosis predictions.

Precision: Precision quantifies the proportion of predicted positive results that are actual positive cases of pneumonia. A precision of 99.87% highlights the model's strong reliability in predicting positive results, greatly reducing the risk of incorrect diagnoses.

F1-score: The harmonic mean of precision and sensitivity, the F1-score provides a balanced measure of performance. The model recorded an impressive F1-score of 99.27%, confirming its balanced efficacy in maintaining a favorable trade-off between precision and recall.

AUC-ROC: The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) measures the model's capability to distinguish between classes across all classification thresholds. An AUC-ROC value of 99.58% underscores the model's excellent overall classification capability, demonstrating its consistency across various threshold settings.

A detailed threshold analysis indicated the optimal classification threshold at 0.90, determined based on balanced accuracy. At this threshold, the model achieved a balanced accuracy of 97.37%, optimizing both sensitivity and specificity for practical clinical deployment.

The confusion matrix further validated these metrics with 1818 true negatives, 84 false positives, 25 false negatives, and 1877 true positives. These results were further visualized through confusion matrix heatmaps, ROC curves, and Precision-Recall curves, providing clear visual insights into the model's discriminative abilities and areas of strength.

These rigorous assessments collectively confirmed the proposed model's high reliability and potential effectiveness as a diagnostic aid for pneumonia detection in clinical environments.

3.6. Libraries and frameworks used for LDCS2

1. The Data ingestion & preprocessing (CPU)

Images are loaded from disk, resized to 224×224, and normalized (ImageNet mean/std) using OpenCV and NumPy.

A PyTorch DataLoader with multiple worker threads handles I/O and batching to keep the GPU fed.

2. Data augmentation (CPU workers)

Random affine transforms, brightness/contrast adjustments, Gaussian noise/blur, grid distortions, and Mixup are applied on the CPU in parallel via DataLoader workers, avoiding GPU contention.

3. Model training & inference (GPU)

The DenseNet-121 backbone plus custom dropout + FC head run on an NVIDIA RTX 4080 (12 GB).

PyTorch v2.1.2+cu118 orchestrates forward/backward passes, with CUDA and cuDNN accelerating convolutions, pooling, and normalization.

The AdamW (AMSGrad) optimizer, ReduceLROnPlateau scheduler, and early-stopping loops all execute within the PyTorch kernel.

4. Evaluation & visualization (CPU)

After a training/inference batch completes, metrics (accuracy, sensitivity, specificity, ROC-AUC) are computed on the CPU.

Matplotlib (and optionally Seaborn) generate ROC/PR curves and confusion-matrix heatmaps without tying up GPU memory.

3.7. User interface of the LDCS2 prototype

The LDCS2 prototype exposes a simple, Bootstrap 5.3.0-based web interface (Figure 2) that was designed explicitly to meet the needs of busy clinical staff and radiographers. At the top, a succinct data-entry form gathers only the most essential patient metadata (name, ID, date of birth, gender), minimizing typing effort and avoiding distraction from clinical tasks. Immediately below, a large drag-and-drop zone accepts DICOM, JPEG, or PNG chest-X-ray files in one step, eliminating the need for cumbersome file-naming conventions or external pre-processing.

Once an image is in place, clicking Analyze Image instantly invokes the DenseNet-121 model in the background. A clear progress indicator and color-coded probability bars (green for "Normal," red for "Pneumonia") give immediate visual feedback, enabling rapid triage decisions. No deep technical knowledge is required – clinicians see only "Pneumonia Probability = 87%" rather than raw logits or thresholds.

Beneath the analysis area, a Recent Predictions panel serves as a lightweight audit log. It lists past patients (or runs) in reverse chronological order, showing thumbnail-X-rays, key metadata, and the model's result at a glance. Quick "Edit" and "Delete" actions let users correct mis-entered patient data or clear out test cases, so the system stays in sync with the electronic health record.

By focusing on simplicity, minimal clicks, and instant, color-driven feedback, this interface directly supports the LDCS2 goals of (1) speeding up pneumonia screening in resource-limited settings, (2) reducing cognitive load for non-AI-savvy staff, and (3) preserving an audit trail for quality control and follow-up.

Figure 2. User interface of the LDCS2 prototype

4. Results and limitations of LDCS2

4.1. Model performance

The DenseNet-121 model was trained for pneumonia detection on a balanced dataset to address class imbalance, achieving robust predictive performance across several evaluation metrics. After training for a total of 26 epochs, early stopping was triggered due to no improvement in validation accuracy for 15 consecutive epochs, indicating convergence of the model.

The highest validation accuracy achieved was 90.74% at epoch 11. The final model selection was based on this highest validation performance. Evaluation on the independent test dataset yielded excellent performance metrics, demonstrating the effectiveness and generalizability of the trained model. Key performance metrics at the optimal classification threshold of 0.90 (selected based on balanced accuracy) were:

Accuracy: 97.13%, Sensitivity (Recall): 98.69%, Specificity: 95.58%, Precision: 99.87%, F1-Score: 99.27%, ROC AUC: 99.58%.

These results were derived from a threshold analysis conducted to optimize the balanced accuracy of the binary classification task. A summary of threshold analysis highlighted a notable trade-off between sensitivity and specificity, with the optimal balance found at a threshold of 0.90.

Disclaimer. Reported accuracies may be inflated due to potential image overlap across the three source datasets, which could have introduced information leakage into the test set.

4.2. Threshold selection analysis

A detailed threshold analysis was conducted to identify the best operating point for classification. Accuracy was observed to range significantly depending on the decision threshold, emphasizing the importance of selecting an appropriate threshold to balance false positives and false negatives in clinical settings.

At a lower threshold (0.10), the model attained 100% sensitivity but had a specificity of just 39.01%, resulting in a lower accuracy of 69.51%.

Incrementally increasing thresholds led to improved specificity and overall accuracy:

Threshold 0.50: Accuracy reached 89.62% with sensitivity remaining at 100% and specificity at 79.23%.

Ultimately, the optimal threshold of 0.90 was selected, providing a balanced accuracy of 97.37% by effectively balancing sensitivity and specificity.

4.3. Visualization of model performance

In order to better understand the results, 6 diagnostic plots were generated:

The presented Receiver Operating Characteristic (ROC) curve, as shown in Figure 3, illustrates the performance of a deep learning model in distinguishing between two classes, normal and pneumonia cases. The ROC curve is a graphical representation of the trade-off between sensitivity (true positive rate) and specificity (false positive rate) across various classification thresholds.

The blue curve represents the model's ability to differentiate between the two classes, with a near-perfect trajectory towards the upper-left corner, indicating high classification performance. The Area Under the Curve (AUC) value is 0.9958, demonstrating the model's exceptional ability to discriminate between positive and negative cases. The diagonal red dashed line represents a random classifier, serving as a baseline reference.

This high AUC score suggests that the model maintains a low false positive rate while achieving a high true positive rate, making it highly effective for pneumonia detection. The curve's steep rise near the origin further indicates strong sensitivity, ensuring that a majority of actual pneumonia cases are correctly identified while keeping misclassifications minimal.

The Precision-Recall (PR) curve, as shown in Figure 4, evaluates the performance of a DenseNet-121 model trained for pneumonia detection. The PR curve illustrates the relationship between precision (positive predictive value) and recall (sensitivity) across various classification thresholds.

The model demonstrates high precision and high recall across most thresholds, as indicated by the curve's sustained elevation near 1.0 until a steep decline towards the far right. The Area Under the PR Curve (AUC-PR) is 0.996, signifying superior model performance, particularly in handling imbalanced datasets where the minority class (pneumonia cases) is of primary interest.

A high AUC-PR score confirms the model's capability in maintaining low false positive rates while capturing the majority of true positives, making it well-suited for medical diagnostic tasks. The near-perfect curve further suggests minimal trade-offs between precision and recall, reinforcing the model's robustness.
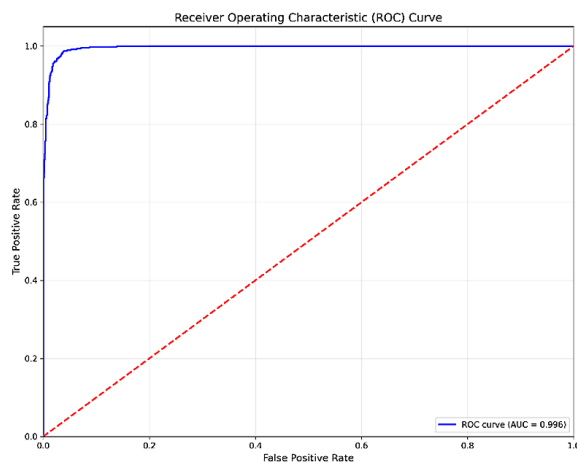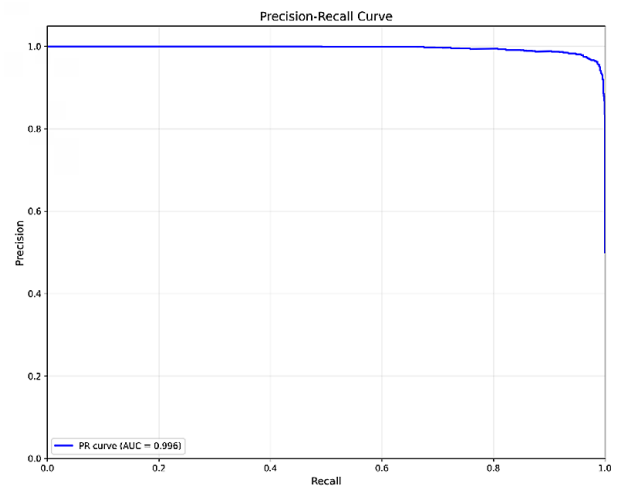

Figure 4. Precision-recall curve

The confusion matrix, as shown in Figure 5, provides a detailed performance evaluation of the DenseNet-121 model trained for pneumonia detection. The matrix consists of four quadrants, indicating the model's classification outcomes:

True Negatives (TN): 1.818 cases were correctly classified as "Normal."

False Positives (FP): 84 normal cases were misclassified as "Pneumonia."

False Negatives (FN): 25 pneumonia cases were misclassified as "Normal."

True Positives (TP): 1.877 cases were correctly classified as "Pneumonia."

The model exhibits high classification performance, with a low false negative rate, suggesting that most pneumonia cases were successfully detected. Furthermore, the false positive rate is minimal, ensuring reliable diagnostic predictions. The high sensitivity (0.9869) and specificity (0.9558), as derived from the confusion matrix, reinforce the model's robustness in distinguishing between normal and pneumonia cases, making it a reliable tool for medical image analysis.
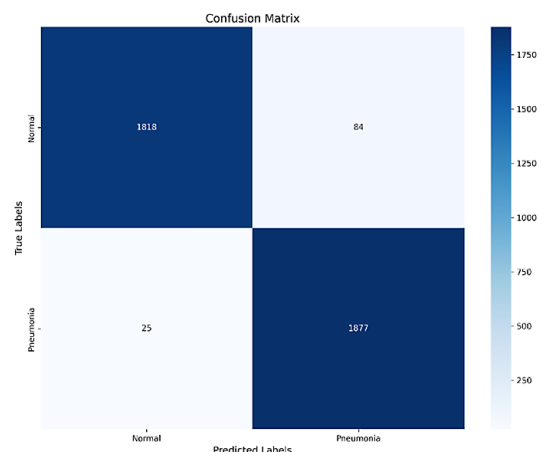

Figure 3. ROC curve

Figure 5. Confusion matrix heatmap

The training and validation curves, as illustrated in Figure 6, provide insight into the learning dynamics of the DenseNet-121 model employed for pneumonia classification. The first plot illustrates the training and validation loss over the course of 26 epochs. The training loss (blue curve) remains consistently low, indicating that the model effectively learns the patterns in the training dataset. However, the validation loss (red curve) exhibits high variance, with notable spikes, suggesting potential overfitting or sensitivity to variations in the validation data.

The second plot depicts the training and validation accuracy across epochs. The training accuracy (blue curve) reaches nearly 100%, further confirming that the model learns well on the training set. In contrast, the validation accuracy (red curve) fluctuates significantly, peaking at 90.74% but experiencing multiple declines. This instability indicates generalization challenges, possibly arising from dataset imbalances or overfitting.

Overall, while the model achieves high training accuracy, the observed disparity in validation performance underscores the need for further regularization, additional data augmentation, or fine-tuning of hyperparameters to enhance generalization to unseen data.

The learning rate schedule, as depicted in Figure 7, illustrates the adaptation of the learning rate throughout training in the pneumonia classification model based on DenseNet-121. The model was trained using the AdamW optimizer with an initial learning rate of 0.00020, and the ReduceLROnPlateau scheduler was employed to dynamically adjust the learning rate in response to validation performance.

The plot demonstrates that the learning rate remained constant during the initial 16 epochs before experiencing the first drop at epoch 17, reducing by a factor of 0.5 to 0.00010. A subsequent reduction occurred at epoch 23, lowering the rate further to 0.00005, indicating that the model may have encountered a plateau in performance. The final decrease took place at epoch 25, stabilizing at
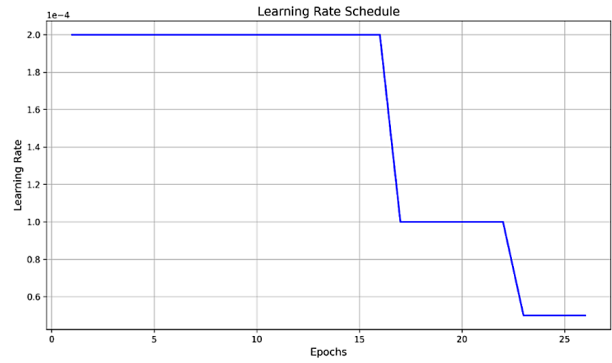

Figure 7. Learning rate schedule

approximately 0.00004 before early stopping was triggered at epoch 26.

This controlled reduction in learning rate suggests an effort to fine-tune model weights more precisely during later stages of training, preventing abrupt weight updates and improving convergence. However, the observed fluctuations in validation performance indicate that additional hyperparameter tuning may be required for optimal generalization.

The epoch training time analysis, as illustrated in Figure 8, demonstrates the computational efficiency of the DenseNet-121 model during training. The plot shows a significant drop in training time after the first epoch, stabilizing at approximately 120–130 seconds per epoch for the remainder of the training process.

The initial high training time can be attributed to several factors, including PyTorch CUDA optimizations, caching of data augmentations, and initialization of model parameters. The subsequent reduction in training time suggests that caching mechanisms and GPU optimizations contributed to improved efficiency. However, slight fluctuations in training duration across later epochs indicate that variations in batch-level computations, memory
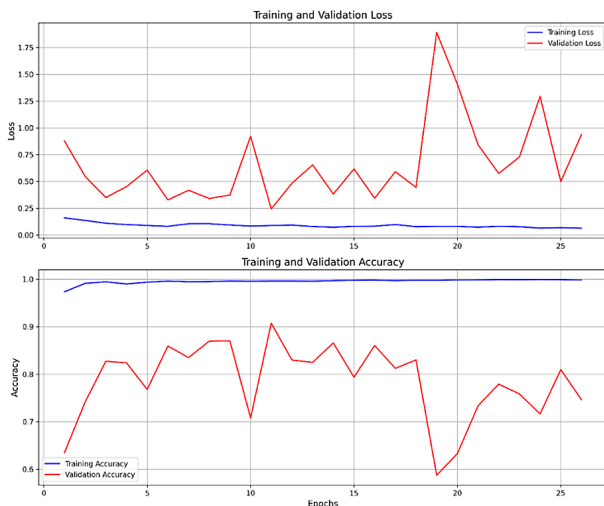

Figure 6. Training and validation loss and accuracy curves
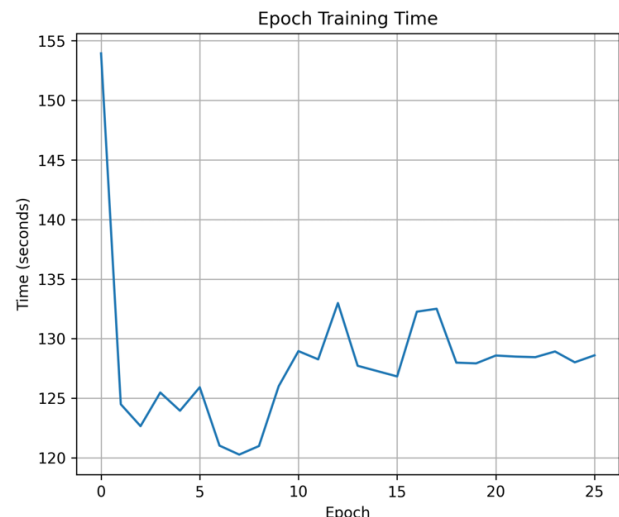

Figure 8. Epoch training time

utilization, and data augmentation complexity influenced training speed.

Overall, the stability in epoch duration demonstrates the computational consistency of the training pipeline. The results confirm that hardware acceleration (NVIDIA RTX 4080 Laptop GPU) was effectively utilized, ensuring efficient deep learning training while maintaining consistent training time across multiple epochs.

These plots facilitated a comprehensive interpretation of model training dynamics, performance trade-offs, and final predictive effectiveness. All visualizations are available and stored for future reference, providing transparency and interpretability for further analysis and clinical review.

In summary, the DenseNet-121 model has shown exceptional predictive capability, high sensitivity, and strong specificity, making it highly suitable for automated pneumonia detection from chest X-rays. Further exploration through Explainable AI (XAI), specifically Grad-CAM visualizations, will enhance model interpretability and clinical trust.

4.4. Computational efficiency

The computational efficiency of the DenseNet-121 model was carefully evaluated to assess its suitability for clinical deployment. Training was conducted using an NVIDIA GeForce RTX 4080 Laptop GPU 60W with a total memory capacity of 12.00 GB. The training process utilized only 4.56 GB of GPU memory per epoch, significantly below the available maximum, which demonstrates efficient use of GPU resources.

Each training epoch lasted on average 128.02 seconds, showcasing rapid training times that are conducive to iterative model development and rapid deployment in clinical settings. The training concluded in approximately 1 hour and 1 minute after 26 epochs due to early stopping triggered by the best validation accuracy reaching 90.74% at epoch 11, ensuring computational resources were conserved without compromising model performance.

The computational efficiency demonstrated in this study, coupled with the minimal GPU memory usage, suggests that the DenseNet-121 model is highly suitable for deployment in clinical environments, even where computational resources are moderately constrained. The use of AdamW optimizer with AMSGrad, a learning rate scheduler, and sophisticated data augmentation strategies contributed significantly to the model's computational and predictive efficiency.

4.5. Limitations

Despite impressive performance metrics, several limitations must be considered. Firstly, the generalizability of the model might be constrained by the characteristics of the datasets used, predominantly comprising pediatric chest radiographs (Mooney, 2018; Sait et al., 2022). Thus, further validation on diverse and adult populations is necessary to ensure broader applicability.

Secondly, the heavy reliance on data augmentation techniques, although beneficial in enhancing generalization, may inadvertently introduce artificial patterns, potentially leading to slight overestimations of model performance in real-world settings (Shorten & Khoshgoftaar, 2019).

Thirdly, the threshold selection process, although optimized for balanced accuracy, might require adjustment when applied to clinical scenarios with different risk tolerances or prevalence rates. The fixed threshold approach could limit the flexibility needed for personalized clinical decision-making.

Fourthly, despite achieving high specificity and sensitivity, the model's interpretability remains limited due to its deep-learning nature, often referred to as a "black-box" model. This lack of interpretability may hinder clinician trust and adoption, necessitating the integration of explainable AI techniques to provide greater insight into model predictions (Rudin, 2019).

Finally, the study aggregated three chest X-ray datasets. Identical or near-identical images may appear in more than one source, creating a possibility that duplicates reached the test partition. Such leakage can yield optimistic performance estimates; the true generalization accuracy may therefore be lower. Future work should incorporate image-level de-duplication and patient-level disjoint splits, and confirm findings on an independent cohort.

4.6. Future directions

Future research directions will focus on expanding the diagnostic capability of the model to include classification of four or more disease categories, addressing the challenges observed in multi-class scenarios. Incorporation of explainable artificial intelligence (XAI) techniques, specifically Grad-CAM, will enhance interpretability and clinical acceptance of the model by elucidating the rationale behind model predictions (Selvaraju et al., 2020). Additional future work includes assessing model performance across diverse demographic groups and investigating integration into clinical workflows to validate its utility as an effective clinical decision-support tool.

5. Conclusions

The DenseNet-121-based deep learning model developed for pneumonia detection demonstrates strong computational performance and predictive accuracy, making it well-suited for practical medical applications. Throughout training, significant emphasis was placed on robustness and reliability, leveraging comprehensive data augmentation strategies, balanced class distributions, and sophisticated training methodologies such as early stopping, Mixup augmentation, and a customized AdamW optimizer with AMSGrad.

The final trained model achieved a best validation accuracy of 90.74% at epoch 11 out of 26 completed epochs. Evaluation on the independent test set demonstrated

impressive performance, achieving an optimal accuracy of 97.13% at a classification threshold of 0.90. This threshold provided a well-balanced trade-off, yielding high sensitivity (98.69%), specificity (95.58%), and precision (99.87%), alongside an F1-score of 99.27%. The ROC-AUC of 99.58% further highlights the model's strong ability to distinguish between normal and pneumonia cases effectively.

Computational efficiency was noteworthy, with the model leveraging GPU acceleration and maintaining modest GPU memory usage (approximately 4.56 GB per epoch). Average epoch completion time was around 128 seconds, demonstrating effective resource utilization.

Future development will enhance the model's clinical relevance by incorporating four or more additional classes, enabling the detection of a broader range of medical conditions. Moreover, to address the critical demand for explainability in clinical deployments, Explainable AI (XAI) techniques such as Grad-CAM will be integrated. Grad-CAM's visualization of model predictions will provide clinicians transparent insights into decision-making processes, bolstering trust and facilitating model validation in medical environments.

In summary, the current implementation represents a robust foundation, poised for further improvements through expanded class coverage and integration of explainability techniques, ultimately enhancing its clinical applicability and interpretability.

References

Abbas, A., Abdelsamea, M. M., & Gaber, M. M. (2020). Classification of COVID-19 in chest X-ray images using DeTraC deep convolutional neural network. *Applied Intelligence*, *51*, 854–864. https://doi.org/10.1007/s10489-020-01829-7

Albahli, S., & Yar, G. N. A. H. (2021). Fast and accurate detection of COVID-19 along with 14 other chest pathologies using a multi-level classification: Algorithm development and validation study. *Journal of Medical Internet Research*, *23*(6), Article e23693. https://doi.org/10.2196/23693

Das, A. K., Ghosh, S., Thunder, S., Dutta, R., Agarwal, S., & Chakrabarti, A. (2021). Automatic COVID-19 detection from X-ray images using ensemble learning with convolutional neural network. *Pattern Analysis and Applications*, *24*, 1111–1124. https://doi.org/10.1007/s10044-021-00970-4

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). *An image is worth 16×16 words: Transformers for image recognition at scale*. ArXiv. https://doi.org/10.48550/arXiv.2010.11929

Elshazly, H., Linse, C., Barth, E., & Martinetz, T. (2020). *Explainable COVID-19 detection using chest X-rays and transfer learning*. ArXiv. https://doi.org/10.48550/arXiv.2012.04833

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., & Liu, Z. (2022). A survey on vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(1), 87–110. https://doi.org/10.1109/TPAMI.2022.3152247

Hassantabar, S., Ahmadi, M., & Sharifi, A. (2020). Diagnosis and detection of infected tissue of COVID-19 patients based on lung x-ray image using convolutional neural network approaches. *Chaos, Solitons & Fractals*, *140*, Article 110170. https://doi.org/10.1016/j.chaos.2020.110170

Ibrahim, D. M., Elshennawy, N. M., & Sarhan, A. M. (2021). Deepchest: Multi-classification deep learning model for diagnosing COVID-19, pneumonia, and lung cancer chest diseases. *Computers in Biology and Medicine*, *132*, Article 104348. https://doi.org/10.1016/j.compbiomed.2021.104348

Irmak, E. (2021). COVID-19 disease severity assessment using CNN model. *IET Image Processing*, *15*(8), 1814–1824. https://doi.org/10.1049/ipr2.12153

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D. A., Halabi, S. S., Sandberg, J. K., Jones, R., Larson, D. B., Langlotz, C. P., Patel, B. N., Lungren, M. P., & Ng, A. Y. (2019). CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(1), 590–597. https://doi.org/10.1609/aaai.v33i01.3301590

Ismael, A. M., & Şengür, A. (2021). Deep learning approaches for COVID-19 detection based on chest X-ray images. *Expert Systems with Applications*, *164*, Article 114054. https://doi.org/10.1016/j.eswa.2020.114054

Katz, S. E., & Williams, D. J. (2018). Pediatric community-acquired pneumonia in the United States: Changing epidemiology, diagnostic and therapeutic challenges, and areas for future research. *Infectious Disease Clinics of North America*, *32*(1), 47–63. https://doi.org/10.1016/j.idc.2017.11.002

Mooney, P. (2018). *Chest X-ray images* (*Pneumonia*). https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia

Moujahid, H., Cherradi, B., El Gannour, O., Bahatti, L., Terrada, O., & Hamida, S. (2020). Convolutional neural network-based classification of pneumonia from chest X-ray images. *African Journal of Science, Technology, Innovation and Development*, *12*(4), 499–506. https://doi.org/10.25046/aj050522

Podder, P., Bhattacharjee, S., & Roy, A. (2021). An efficient method of detection of COVID-19 using Mask R-CNN on chest X-ray images. *AIMS Biophysics*, *8*(3), 281–290. https://doi.org/10.3934/biophy.2021022

Rahman, M. M., Nooruddin, S., Hasan, K. M. A., & Dey, N. K. (2021a). HOG + CNN Net: Diagnosing COVID-19 and pneumonia by deep neural network from chest X-ray images. *SN Computer Science*, *2*, Article 371. https://doi.org/10.1007/s42979-021-00762-x

Rahman, T., Khandakar, A., Qiblawey, Y., Tahir, A., Kiranyaz, S., Kashem, S. I., Chowdhury, M. E. H., Al-Maadeed, S., Zughaier, S. M., & Khan, M. S. (2021). Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images. *Computers in Biology and Medicine*, *132*, Article 104319. https://doi.org/10.1016/j.compbiomed.2021.104319

Rajpurkar, P., Irvin, J., Ball, R. L., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D. Y., Bagul, A., Langlotz, C. P., Patel, B. N., Yeung, S. Y., & Ng, A. Y. (2018). Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Medicine*, *15*(11), Article e1002686. https://doi.org/10.1371/journal.pmed.1002686

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D. Y., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M. P., & Ng, A. Y. (2017). *CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning*. ArXiv. https://doi.org/10.48550/arXiv.1711.05225

Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, T., Ursprung, S., Aviles-Rivero, A. I., Etmann, C., McCague, C., Beer, L., Weir-McCall, J. R., Teng, Z., Gkrania-Klotsas, E., AIX-COVNET, Rudd, J. H. F., Sala, E., & Schönlieb, C.-B. (2021). Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs

and CT scans. *Nature Machine Intelligence*, *3*(3), 199–217. https://doi.org/10.1038/s42256-021-00307-0

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*, 206–215. https://doi.org/10.1038/s42256-019-0048-x

Sahiner, B., Pezeshk, A., Hadjiiski, L. M., Wang, X., Drukker, K., Cha, K. H., Summers, R. M., & Giger, M. L. (2019). Deep learning in medical imaging and radiation therapy. *Medical Physics*, *46*(1), e1–e36. https://doi.org/10.1002/mp.13264

Sait, U., KV, G. L., Prajapati, S. P., Bhaumik, R., Kumar, T., Shivakumar, S., & Bhalla, K. (2022). *Curated dataset for COVID-19 posterior-anterior chest radiography images* (*X-Rays*). https://doi.org/10.17632/9xkhgts2s6.4

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 618–626). IEEE. https://doi.org/10.1109/ICCV.2017.74

Shamshad, F., Khan, S., Zamir, S. W., Khan, M. H., Hayat, M., Khan, F. S., & Fu, H. (2023). Transformers in medical imaging: A survey. *Medical Image Analysis*, *88*, Article 102802. https://doi.org/10.1016/j.media.2023.102802

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, *6*(1), Article 60. https://doi.org/10.1186/s40537-019-0197-0

Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. *Proceedings of the International Conference on Machine Learning* (pp. 6105–6114). https://doi.org/10.48550/arXiv.1905.11946

Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PLoS ONE*, *14*(11), Article e0224365. https://doi.org/10.1371/journal.pone.0224365

Varshni, D., Thakral, K., Agarwal, L., Nijhawan, R., & Mittal, A. (2019). Pneumonia detection using CNN based feature extraction. In *2019 IEEE International Conference on Electrical, Computer and Communication Technologies* (*ICECCT*) (pp.1–7). IEEE. https://doi.org/10.1109/ICECCT.2019.8869364

Wang, B., & Zhang, W. (2022). MARnet: Multi-scale adaptive residual neural network for chest X-ray images recognition of lung diseases. *Mathematical Biosciences and Engineering*, *19*(1), 331–350. https://doi.org/10.3934/mbe.2022017

World Health Organization. (2022). *Pneumonia*. https://www.who.int/news-room/fact-sheets/detail/pneumonia

PNEUMONIJOS NUSTATYMAS IŠ KRŪTINĖS LĄSTOS RENTGENOGRAMŲ, NAUDOJANT KONVOLIUCINIUS NEURONINIUS TINKLUS

P. Bundza, J. Trinkūnas

Santrauka

Pneumonijos nustatymas iš krūtinės ląstos rentgenogramų yra itin svarbus ankstyvajai diagnostikai, o giliojo mokymosi modeliai – ypač konvoliuciniai neuroniniai tinklai (CNN) – rodo didelį potencialą automatizuojant šį procesą. Šiame tyrime sukurtas ir apmokytas CNN, paremtas *DenseNet-121* architektūra ir pavadintas LDCS2, skirtas klasifikuoti krūtinės ląstos rentgeno vaizdams, iš kurių matyti pneumonija arba sveiki plaučiai, naudojant sujungtą duomenų rinkinį iš trijų viešai prieinamų šaltinių. CNN metodas pasirinktas vietoje *Vision Transformers* (ViT) dėl mažesnių skaičiavimo išteklių reikalavimų ir geresnių rezultatų, kai duomenų kiekis ribotas. Siekiant sutrumpinti vykdymo laiką, vietoje *k* kartų kryžminės validacijos taikytas tradicinis mokymo, validacijos ir testavimo skaidymas. LDCS2 pademonstravo puikią atskyrimo gebą tarp pneumonijos ir sveikų plaučių vaizdų bei aukštą skaičiavimo efektyvumą. Šie rezultatai pabrėžia *DenseNet* pagrindu veikiančių CNN potencialą automatizuotai plaučių uždegimo diagnostikai, ypač išteklių stokojančiose aplinkose.

Reikšminiai žodžiai: LDCS2, konvoliuciniai neuroniniai tinklai (CNN), krūtinės ląstos rentgeno vaizdų klasifikavimas, pneumonijos aptikimas, *DenseNet-121*, medicininis vaizdavimas, gilusis mokymasis sveikatos priežiūros srityje, duomenų augmentacija, perkėliminis mokymasis.