Jurnal Informatika: Jurnal pengembangan IT

Vol. 11, No. 2, 2026
ISSN : 2477-5126, DOI:10.30591/jpit.v11i2.10374

Optimizing Pneumonia Detection on Edge Devices Using YOLOv5lu with Pruning and Quantization

Muhammad Satriya Pratama Manggala Kusuma [1], Cinantya Paramita [2], Suryanti [3]
[1,2] Faculty of Computer Science, Universitas Dian Nuswantoro, Indonesia
[2] Dinus Research Group for Al in Medical Science (DREAMS), Universitas Dian Nuswantoro, Indonesia
[3] Faculty of Medicine, Universitas Dian Nuswantoro, Indonesia
111202214225@mhs.dinus.ac.id, cinantya.paramita@dsn.dinus.ac.id, suryanti@dsn.dinus.ac.id

Info Artikel

Riwayat Artikel:

Received 2026-03-02
Revised 2026-07-01
Accepted 2026-07-19

Abstract **–** Pneumonia continues to be a leading cause of childhood mortality worldwide, with the greatest impact in resource limited regions where access to radiologists is limited. Deep learning has emerged as a promising tool for automated screening, yet conventional object detection models demand high computational resources, restricting their use on low cost edge devices common in rural healthcare settings. To address this challenge, we developed a lightweight pneumonia detection framework designed for realtime inference on standard hardware. Our approach builds on the YOLOv5lu architecture, which incorporates an anchor free decoupled head, and was trained on a balanced dataset of 14,863 chest X-rays from the RSNA Pneumonia Detection Challenge.To enable deployment on edge devices, we applied a sequential compression pipeline. First, 30% of convolutional filters were removed through structured pruning, followed by post training quantization to INT8. These optimizations reduced the model size by 48.79% (from 101 MB to 51.72 MB) and achieved a 36.46% improvement in inference speed, accelerating performance to 6.24 ms per image, equivalent to more than 160 frames per second on a standard CPU. Importantly, the quantized model preserved diagnostic performance, achieving a mean Average Precision (mAP@0.4–0.75) of 0.266 compared to the baseline score of 0.287. These findings confirm the practical feasibility of deploying advanced deep learning models in limited resource regions. By effectively balancing efficiency and accuracy, this framework offers a scalable solution for early pneumonia screening and establishes a foundation for extending detection to other diseases, including Tuberculosis and COVID-19.

Keywords: *Model Compression; Medical Imaging; Pneumonia Detection; YOLOv5.*

Corresponding Author:

Cinantya Paramita

Email:

cinantya.paramita@dsn.dinus.ac.id

This is an open access article under the CC BY 4.0 license.

Abstrak *– Pneumonia masih menjadi salah satu penyebab utama kematian pada anak di seluruh dunia, terutama di wilayah dengan keterbatasan sumber daya dan minimnya akses terhadap tenaga radiolog. Pemanfaatan deep learning untuk skrining otomatis menunjukkan potensi besar, namun sebagian besar model deteksi objek konvensional membutuhkan sumber daya komputasi yang tinggi sehingga kurang sesuai untuk diterapkan pada perangkat edge berbiaya rendah di fasilitas kesehatan pedesaan. Penelitian ini mengusulkan sebuah kerangka kerja deteksi pneumonia yang ringan dan dioptimalkan untuk inferensi real-time pada perangkat keras standar. Model dibangun berdasarkan arsitektur YOLOv5lu yang mengadopsi anchor-free decoupled head dan dilatih menggunakan dataset seimbang berjumlah 14.863 citra X-ray dada dari RSNA Pneumonia Detection Challenge. Untuk mendukung implementasi pada perangkat edge, diterapkan pipeline kompresi berurutan yang meliputi structured pruning sebesar 30% pada filter konvolusi dan post-training quantization ke format INT8. Hasil optimasi menunjukkan bahwa ukuran model berhasil diperkecil sebesar 48,79% (dari 101 MB menjadi 51,72 MB), serta meningkatkan kecepatan inferensi sebesar 36,46% menjadi 6,24 ms per citra atau lebih dari 160 frame per detik pada CPU standar. Meskipun mengalami proses kompresi, model terkuantisasi tetap mempertahankan kinerja deteksi yang baik dengan nilai mean Average Precision (mAP@0,4–0,75) sebesar 0,266, dibandingkan dengan skor baseline sebesar 0,287. Hasil penelitian ini menunjukkan bahwa model deep learning yang efisien secara komputasi dapat diterapkan secara praktis di lingkungan dengan sumber daya terbatas. Kerangka kerja yang diusulkan berpotensi menjadi solusi yang skalabel untuk skrining dini pneumonia serta dapat dikembangkan lebih lanjut untuk deteksi penyakit lain seperti Tuberkulosis dan COVID-19.*

Kata Kunci*: Deteksi Pneumonia, Kompresi Model, Pencitraan Medis, YOLOv5.*

I. INTRODUCTION

Pneumonia remains the leading cause of infectious death among children under five years of age globally, and since 2010 progress to reduce these deaths has lagged behind diarrhea, measles and malaria[1]. According to the World Health Organization (WHO) this disease is responsible for around 15% of total infant deaths, leading to more than 740,000 fatalities each year in children younger than five[2]. In 2022 alone, pneumonia led to over 1,231,000 emergency department visits in the United States[3] and caused more than 41,000 deaths[4], ranking it among the nation's top ten causes of death. This disorder drastically reduces the capability of those affected to breathe properly and absorb sufficient oxygen into their blood. The World Health Organization (WHO) highlights pneumonia as a major worldwide health issue[5]. The severity and widespread impact of this disease underscore the urgent need for

early and accurate detection methods to facilitate prompt treatment and management, thereby reducing its substantial public health burden[5], [6]. Chest X-rays are commonly used to detect pneumonia by showing lung opacity and consolidation, however, manual interpretation is inherently subjective and prone to significant inter-observer variability, often complicated by radiological similarities with other pulmonary conditions. Quick and accurate diagnosis is vital for effective treatment. In recent years, machine learning models have been developed to help doctors identify pneumonia from chest X-ray images[7].

In resource-limited settings where access to skilled radiologists is often restricted, the challenge of timely and accurate pneumonia diagnosis becomes even more pronounced[8].The Radiological Society of North America (RSNA) Pneumonia Detection Challenge, established in 2018, provides an expertly annotated collection of diverse chest X-rays with the DICOM format labeled pneumonia instances that has facilitated extensive research in machine learning-driven pneumonia detection and enabled comparative evaluation of algorithmic approaches[9]. DICOM (Digital Imaging and Communications in Medicine) is an internationally adopted standard that specifies a non-proprietary format for the storage, transmission, and exchange of medical images and associated metadata across diverse imaging modalities and healthcare institutions[10]. Deep learning based object detection methods have demonstrated outstanding accuracy and speed in real-time detection and localization tasks[11], [12], [13]. Reliable and efficient detection models are crucial in supporting healthcare professionals with diagnosis and treatment of medical conditions[14], [15]. Among these methods, the You Only Look Once (YOLO) algorithm has received considerable recognition[16], [17], [18], [19].

YOLO is a state-of-the-art, object detection algorithm widely recognized in the computer vision field[20]. As a single-stage detector, it identifies all objects in an image through one forward pass of a convolutional neural network (CNN). Using a single network to predict both bounding boxes and class probabilities[21], [22]. However, the substantial computational complexity and memory demands of these deep learning architectures pose a major challenge for deployment on resource-constrained edge devices, thereby necessitating efficient optimization strategies. Model compression techniques including pruning, quantization, and knowledge distillation have become essential for deploying sophisticated deep learning models on resource-constrained edge devices. Pruning operates by systematically removing low-weight parameters and redundant connections, leading to a substantial decrease in both memory requirements and computational costs while maintaining model accuracy[23]. Quantization reduces the precision of model weights by representing parameters with lower-bit precision (e.g., 8bit or 4bit integers instead of 16bit or 32bit floats), which can reduce memory overhead by a factor of 4 and computational cost by a factor of 16 [23]. The integration of pruning and quantization represents a complementary approach where pruning reduces network parameters through connection elimination while quantization further reduces memory footprint through precision reduction, achieving greater compression ratios than either technique alone[24].

While YOLOv5 introduces a versatile and scalable architecture with several innovations aimed at advancing real-time object detection by leveraging flexible backbone networks, efficient training methodologies, and optimized inference strategies[24], deep learning models usually require a lot of computing power, meaning they often cannot run on the low-power machines available in rural clinics. To circumvent this, recent research has explored architectural optimizations for YOLO models. However, a significant research gap remains in prior literature regarding medical imaging object detection on the edge. Previous studies heavily favor anchor-based architectures (such as standard YOLOv5 or YOLOv7), which rely on rigid predefined bounding box templates. While anchor boxes perform adequately for uniform, hard-edged objects in standard computer vision tasks (e.g., pedestrian or vehicle detection), they struggle significantly with the highly irregular, diffuse, and ambiguous pathological boundaries typical of lung opacities in medical imaging. Furthermore, attempts to apply compression techniques to medical object detectors often employ hardware-incompatible or overly complex optimization pipelines that cause severe drops in clinical sensitivity. For instance, recent works utilizing structured pruning and distillation on aerial datasets or standard human detection tasks do not account for the fragile feature mapping required to identify low-contrast clinical abnormalities[25]. In the medical imaging domain, optimization frameworks have primarily focused on standard classification or segmentation rather than the localized regression needed for automated screening pipelines on low-cost hardware.

To bridge these gaps, this study introduces a highly optimized, edge-deployable pneumonia screening framework. The core novelty of this research lies in the unique combination of the YOLOv5-Large Ultralytics (YOLOv5lu) architecture which explicitly leverages an anchor-free decoupled head to eliminate the rigid bias of traditional anchor boxes when mapping amorphous lung lesions with a unified, sequential hardware-friendly compression pipeline consisting of 30% L2-norm structured channel pruning and MinMax INT8 post-training quantization. Consequently, this study establishes a scalable, clinically viable baseline for deploying advanced deep learning models in limited-resource regions without sacrificing localization flexibility.

II. METHODOLOGY

The research workflow used in this study, illustrated in Figure 1, follows a linear sequence to maintain clinical relevance and practical feasibility. The work starts with preparing the dataset and performing basic preprocessing, then continues to the modeling stage using the YOLOv5lu architecture. Since the goal is to make the system usable in healthcare settings with limited resources, the trained model is later optimized through structured pruning and quantization. The process ends with a detailed evaluation to check how accurate the model is and how efficiently it runs during inference.

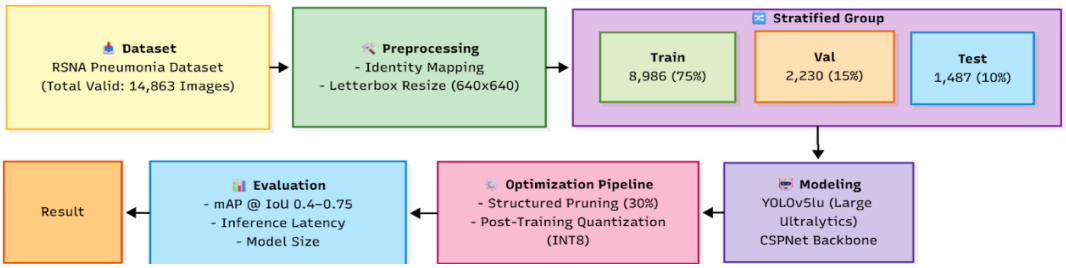


Figure 1. Overall Research Framework

A. *Dataset and Subject Selection*

The foundation of this study is the dataset from the RSNA Pneumonia Detection Challenge (2018), which provides expert-annotated chest X-ray images.

Figure 2. CXR Pneumonia and Normal

To make the model more clinically focused, the original three-class setup was simplified into two classes. The ambiguous “No Lung Opacity/Not Normal” category was removed so that the model only distinguishes clear pathological cases labeled “Lung Opacity” from healthy controls labeled “Normal.” After this curation process, shown in Figure 3, the final dataset contained 14,863 valid patient records (Positive: 6,012; Negative: 8,851).

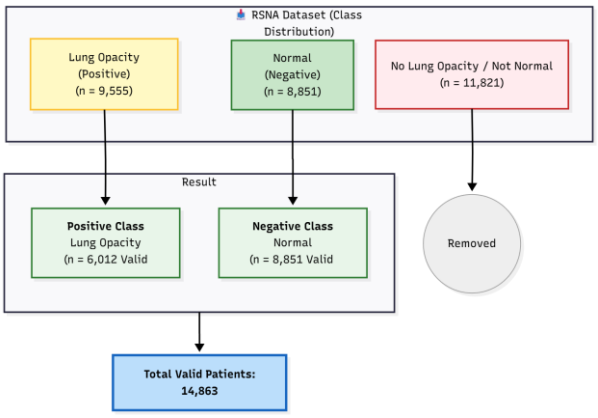


Figure 3. Dataset Distribution

B. *Dataset Partitioning and Stratification Strategy*

To evaluate the model properly and avoid data leakage, we used a patient-level stratified splitting strategy (Figure 4). By applying StratifiedGroupKFold, all images belonging to the same patient were kept within the same subset.

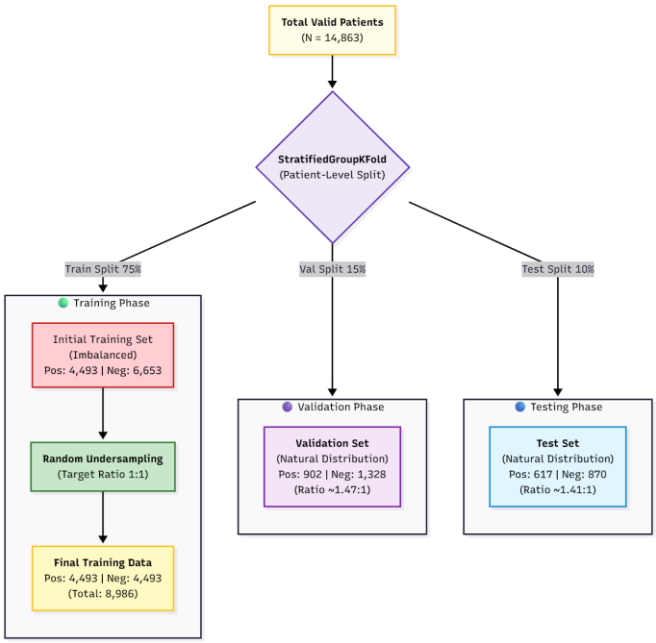


Figure 4. Data Partitioning and Class Balancing

The dataset was divided into three distinct parts. The training set initially consisted of 4,493 positive patients and 6,653 negative patients. To address the class imbalance, random undersampling was applied to the negative class, resulting in a balanced 1:1 ratio and a final training size of 8,986 patients. The validation set included 902 positive and 1,328 negative patients, while the test set comprised 617 positive and 870 negative patients. The validation and test sets retained their original class ratios (around 1.47:1 and 1.41:1, respectively) to reflect real world prevalence and to assess the model's performance under realistic clinical conditions.

C. Image Preprocessing Pipeline

Before being used by the model, the raw DICOM files had to be standardized to make sure the feature extraction process was consistent. As shown in Figure 5, the preprocessing steps convert the high bit depth medical images into a format suitable for deep learning. This includes rescaling pixel intensities based on the DICOM metadata, normalizing the values to the 0–255 range, and correcting the photometric interpretation so that bone structures always appear white.

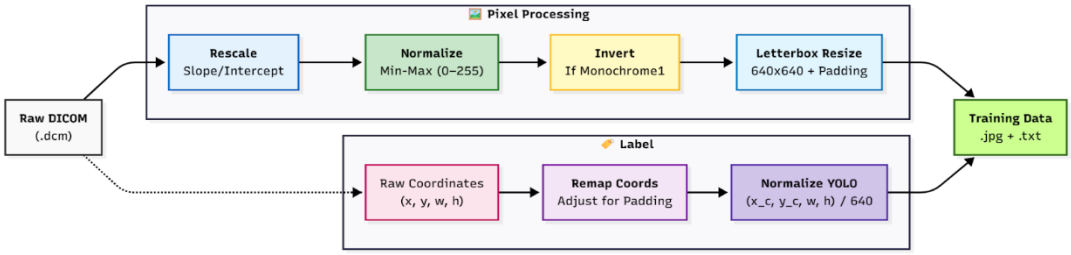


Figure 5. Image Pre-processing Pipeline

All images were then resized to 640 × 640 pixels using letterboxing, which adds padding to preserve the original anatomical aspect ratio and prevents distortions that could mislead the detector. The bounding box coordinates were also adjusted to match the normalized YOLO format:

$$x_{norm} = \frac{x_c}{W}, \quad y_{norm} = \frac{y_c}{H}, \quad w_{norm} = \frac{w}{W}, \quad h_{norm} = \frac{h}{H} \qquad (1)$$

Where (x_c, y_c) is the box center, (w, h) are box dimensions, and (W, H) are image dimensions.

D. Object Detection Architecture

With the data prepared, we used the YOLOv5-Large Ultralytics (YOLOv5lu) architecture as the base detector (Figure 6). This version was chosen because its Decoupled Head design processes class probabilities and bounding box regression separately, allowing it to converge more reliably than earlier anchor-based models. The model was initialized with COCO pretrained weights and trained on an NVIDIA RTX 5070 Ti GPU using Stochastic Gradient Descent (SGD) as the optimizer.

To ensure stable convergence and optimal feature extraction, the training hyperparameters were meticulously configured. The model was trained with a batch size of 16 for a maximum of 100 epochs. The optimization process utilized an initial learning rate of 0.01, which was systematically decayed using a cosine learning rate scheduler down to a final value of 0.001 to prevent gradient oscillations in later stages. The SGD optimizer was further configured with a momentum of 0.937 and a weight decay coefficient of 0.0005 to penalize overly complex weights. To reduce the risk of overfitting on the medical images, we applied a label smoothing coefficient of 0.1 and utilized mosaic augmentation. Training was stopped early via an early stopping callback whenever the validation mAP failed to improve for 20 consecutive epochs, preventing unnecessary computational expenditure and overfitting

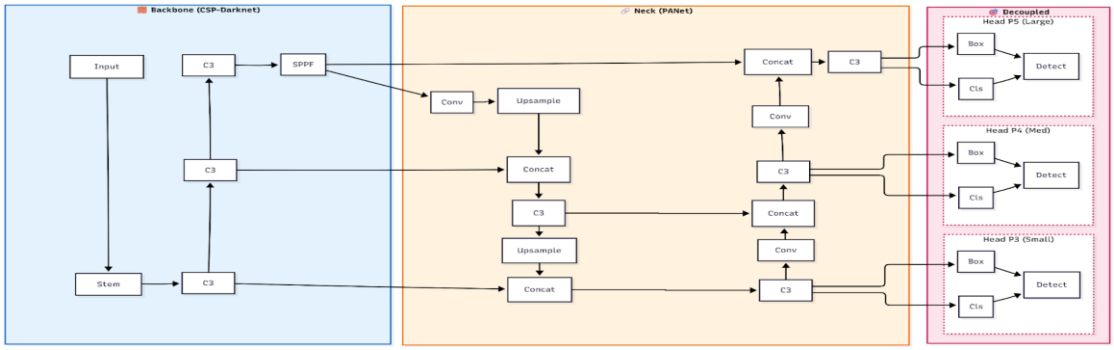


Figure 6. YOLOv5lu Architecture

E. Model Optimization and Compression

Although the YOLOv5lu baseline achieved good accuracy, its computational process can be too heavy for devices in rural clinics. To address this, we implemented an optimization workflow (Figure 7) that first applies structured pruning, followed by quantization.

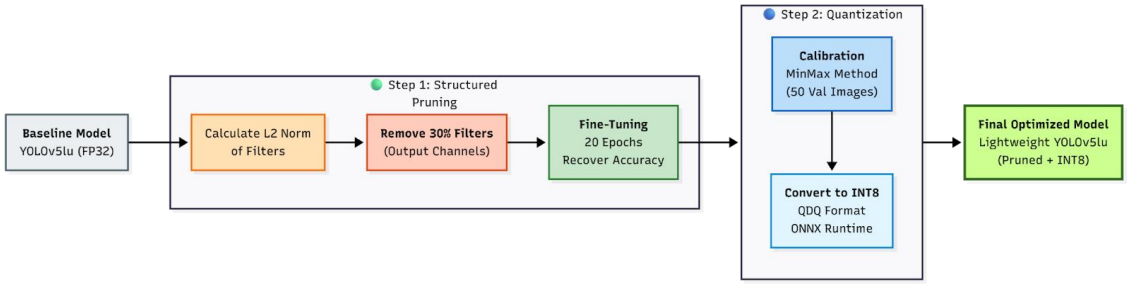


Figure 7. Optimization Pipeline

1) *Structured Pruning*: We applied L2-norm-based structured pruning to reduce the number of model parameters. This method removes entire filters (output channels) based on their magnitude, as filters with smaller weights are considered less important for the network's expressive power. We pruned exactly 30% of the filters across all Conv2d layers. The choice of a 30% pruning threshold represents a well-established empirical baseline in deep neural network compression; empirical literature indicates that pruning up to 30% of convolutional channels typically strips away redundant, low-magnitude feature extractors without disrupting the essential structural topography or causing catastrophic accuracy degradation[23], [24]. After pruning, the model underwent fine-tuning for 20 epochs using an adjusted learning rate of 0.001 to recover any potential drop in accuracy caused by the removed connections.

2) *Post-Training Quantization (PTQ)*: The pruned and fine-tuned model was then compressed using static quantization through ONNX (Open Neural Network Exchange) Runtime. We employed the Quantize-Dequantize (QDQ) format with INT8 precision. Calibration was performed using the MinMax method on a subset of 50 validation images to determine the optimal dynamic range for activation quantization. A calibration size of 50 images was selected because it provides a statistically representative distribution of the lung opacity and normal clinical variations required to compute stable scaling factors and zero-point offsets. In post-training quantization pipelines, utilizing 30 to 50 well-stratified calibration samples strikes an optimal balance: it prevents the out-of-memory errors and data

leaks associated with larger calibration pools while mitigating the severe quantization noise and accuracy drops that occur when using fewer than 20 images. This process reduces the weight precision from 32-bit floating point (FP32) to 8-bit integer (INT8), significantly lowering memory usage.

F. Performance Evaluation Metrics

To validate the efficacy of our proposed framework, we adhered to the specific evaluation protocol of the RSNA Pneumonia Detection Challenge. Performance is assessed using the mean Average Precision (mAP) calculated across a specific range of Intersection over Union (IoU) thresholds. The Intersection over Union (IoU) between a predicted bounding box (A) and a ground truth box (B) is defined as:

$$IoU(A,B) = \frac{A \cap B}{A \cup B} \qquad (2)$$

The evaluation metric sweeps over a range of IoU thresholds from 0.4 to 0.75 with a step size of 0.05 (i.e.,$t \in \{0.40, 0.45, 0.50, 0.55, 0.60, 0.65, 0.70, 0.75\}$). At each threshold t, the precision is calculated based on the number of True Positives (TP), False Negatives (FN), and False Positives (FP):

$$Precision(t) = \frac{TP(t)}{TP(t)+FP(t)+FN(t)} \qquad (3)$$

A prediction is considered a True Positive if its IoU with a ground truth object exceeds the threshold t. A False Positive occurs if a predicted object has no associated ground truth object, or if the IoU is below the threshold.

Note on Negative Samples: If an image contains no ground truth objects (negative case), any number of predicted bounding boxes (False Positives) results in a precision score of zero for that image, penalizing false alarms in healthy patients. The final score for a single image is calculated as the mean of the precision values across all thresholds:

$$AP_{Image} = \frac{1}{|thresholds|}\sum_t \frac{TP(t)}{TP(t)+FP(t)+FN(t)} \qquad (4)$$

In addition to the primary RSNA metric, we report inference latency (ms) to provide a comprehensive analysis of model efficiency suitable for edge deployment.

III. RESULT AND DISCUSSIONS

A. Result

The performance of the proposed pneumonia detection framework was evaluated on the reserved independent test set comprising 1,487 patients. The evaluation focuses on three key dimensions: diagnostic accuracy (mAP), computational efficiency (inference latency), and storage requirements (model size).

1) Diagnostic Performance of the Baseline Model: The model achieved a mAP of 49.53% at the PASCAL VOC threshold (IoU = 0.50). Under the more stringent COCO-style metric (mAP@0.5:0.95), performance was 22.96%, while at IoU = 0.4-0.75 the mAP remained 28.38%. Class-level evaluation yielded a mean Precision of 0.5832, Recall of 0.4636, and an F1-score of 0.5166, reflecting a balanced trade-off between minimizing false positives and maintaining sensitivity to lung opacities. The inference speed analysis highlights the model's efficiency, with a total processing time of 9.82 ms per image, confirming the model's suitability for real-time screening applications.

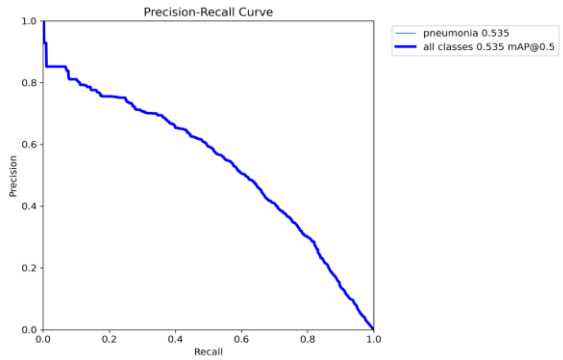


Figure 8. Precision-Recall Curve for the Baseline YOLOv5lu Model.

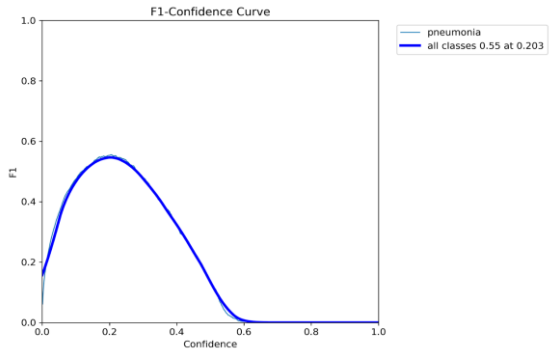


Figure 9. F1-Score vs Confidence Threshold.

2) Impact of Model Optimization Techniques: Applying structured pruning followed by INT8 quantization produced notable improvements in computational efficiency while preserving diagnostic accuracy. Table 1 presents a comparative analysis of the Baseline (FP32), Pruned (FP32), and Pruned + Quantized (INT8) models, highlighting the tradeoffs involved in the optimization pipeline.

TABLE 1
COMPARISON OF MODEL ACCURACY ACROSS OPTIMIZATION STAGES AT DIFFERENT IOU

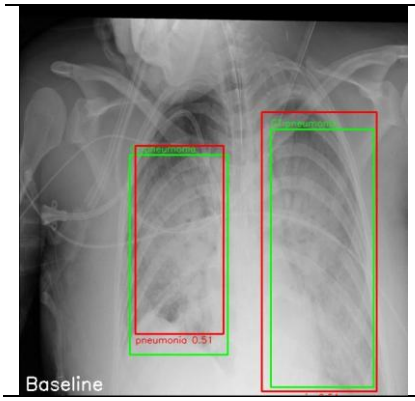
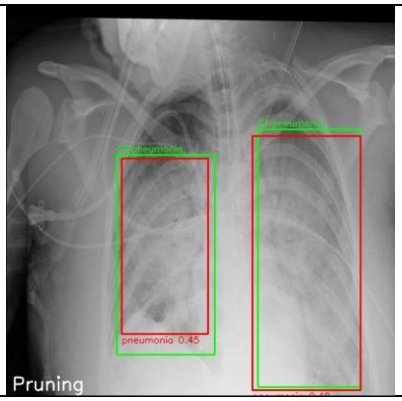
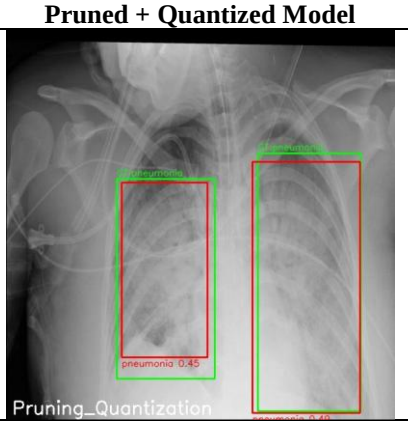
Model	IoU 0.4	IoU 0.45	IoU 0.5	IoU 0.55	IoU 0.6	IoU 0.65	IoU 0.7	IoU 0.75	mAP
Baseline (YOLOv5lu FP32)	0.442	0.419	0.343	0.321	0.253	0.232	0.162	0.120	0.287
Pruned Model (FP32)	0.422	0.427	0.325	0.299	0.261	0.214	0.150	0.123	0.278
Pruned + Quantized (INT8)	0.402	0.431	0.313	0.293	0.263	0.182	0.116	0.125	0.266

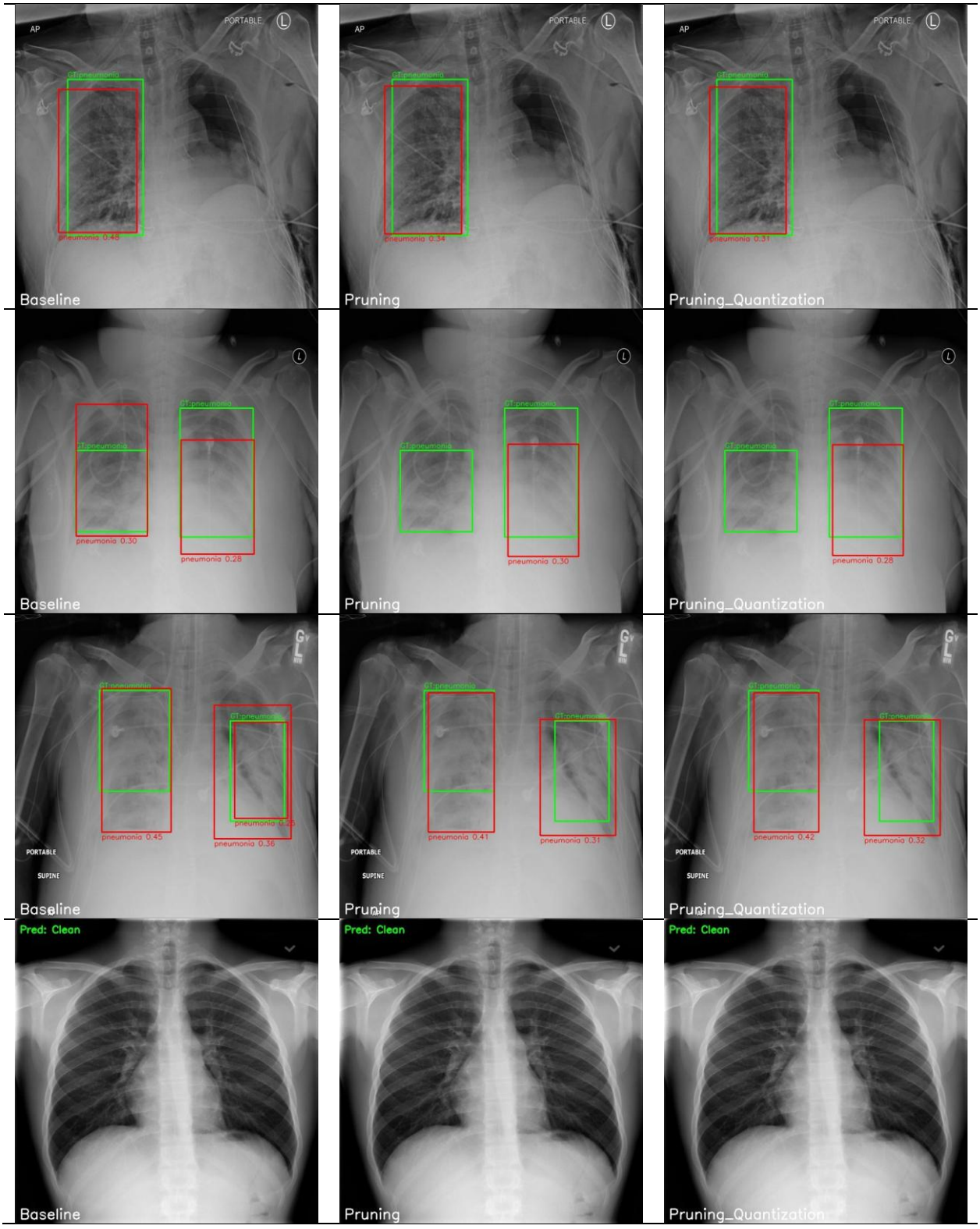
TABLE 2
COMPARISON OF MODEL SIZE, SPEED, AND ACCURACY ACROSS OPTIMIZATION STAGES

Model	Size (MB)	Inference Speed (ms/img)	mAP (IoU 0.4-0.75)
Baseline (YOLOv5lu FP32)	101	9.82ms	0.287
Pruned Model (FP32)	101	8.95ms	0.278
Pruned + Quantized (INT8)	51.72	6.24ms	0.266

3) Qualitative Analysis: To give a clear visual comparison of how each model performs across the optimization pipeline, we created a grid of test samples. For the Pneumonia class, we chose cases where all three model variants (Baseline, Pruned, and Quantized) successfully detected the ground-truth pathology. This allows a fair comparison of bounding box accuracy and confidence scores on the exact same anatomical regions. For the Normal class, we randomly selected samples to check whether the models consistently avoided false positives in healthy lungs.Table 3 shows the results. Green bounding boxes indicate the ground truth annotations, while red boxes show the model predictions. The labels above each detection specify which model variant produced the prediction along with its confidence score.

TABLE 3
QUALITATIVE COMPARISON OF DETECTION RESULTS. ROWS 1-4 DISPLAY PNEUMONIA CASES, WHILE ROW 5 DISPLAYS A NORMAL CONTROL CASE. GREEN BOXES DENOTE GROUND TRUTH; RED BOXES DENOTE MODEL PREDICTIONS WITH CONFIDENCE SCORES.

Baseline Model	Pruned Model	Pruned + Quantized Model
		



B. Discussions

This work shows that deep learning models can be optimized for use in healthcare environments with limited computational resources without experiencing a significant decline in diagnostic accuracy. The sections below explain how these findings relate to the model's design choices, the optimization methods used, and their relevance in clinical practice.

1) Interpretation of Diagnostic Accuracy: The base model YOLOv5lu achieved a competitive mAP score of 0.287 on the RSNA test dataset, a result largely attributable to the innovation of the "u" series architecture. Unlike previous anchor based iterations (YOLOv3, YOLOv4, and standard YOLOv5), YOLOv5lu uses a separate head without anchors. In pneumonia detection, pathologies vary significantly in shape, aspect ratio, and size ranging from from small, localized to extensive infiltrates that can involve entire lobes. Anchor based models often struggle with this irregularity if the predefined anchor boxes do not match the statistics of the dataset. By directly predicting the center of the object, the YOLOv5lu architecture eliminates this bias, enabling more flexible and accurate localization for various shapes of opacities. Moreover, applying class balancing through random undersampling during training was essential. The original RSNA dataset is heavily imbalanced, with normal cases dominating the distribution. Without balancing, object detectors often simplify their results into trivial solutions that predict 'Normal' for all inputs in order to reduce losses. By applying a balanced sampling, the model is required to learn features that distinguish opacity regions, helping improve the recall scores.

2) Empirical Analysis of the Low mAP and Failure Cases: The low mAP baseline and its marginal post-optimization drop stem from the faint, low-contrast nature of pneumonia, which lacks the sharp boundaries found in standard computer vision objects. Qualitative analysis in Table 3 (Row 3) demonstrates that while the model correctly localizes the clinical pathology, subtle bounding box regression variances or coordinate over-extensions lower the strict Intersection over Union (IoU) scores. Under optimization, minor feature sensitivity drops cause the network to merge or omit adjacent, soft-edged opacity regions (Table 3, Row 3). Because the RSNA metric averages precision across a strict threshold range (0.40–0.75), these minor boundary shifts heavily penalize the final mAP, despite successfully alerting clinicians to the disease's presence.

3) Accuracy, Size, and Speed Trade-off: The pipeline exhibits a highly efficient optimization trade-off. Transitioning from the full-precision baseline to the 30% structured pruned variant maintained the 101 MB footprint due to structural zero-padding but lowered latency by 8.85% (from 9.82 ms to 8.95 ms) at a negligible accuracy cost of 0.009 mAP. Subsequent INT8 quantization collapsed the storage footprint by 48.79% down to 51.72 MB and cut inference latency to 6.24 ms per image (a 36.46% total speedup). This massive efficiency gain costs a minor cumulative drop of 0.021 mAP from the baseline, confirming that low-bit integer conversion preserves primary classification features while only marginally blunting sub-pixel boundary regression.

4) Comparative Analysis with Prior Literature: Heavy, uncompressed networks achieve marginal localization gains on chest radiographs but demand massive footprints that cause prohibitive execution bottlenecks on standard edge CPUs. Conversely, our selection of a 30% pruning threshold is strongly supported by the PruneCXR study [27], which established that thoracic deep learning models maintain high structural redundancy and can tolerate up to 60% channel pruning before suffering severe performance drops. By implementing a conservative 30% pipeline, our framework strategically capitalizes on this structural redundancy to eliminate unnecessary filters without pushing the network near its critical degradation boundaries. While PruneCXR warns that pruning typically degrades accuracy in long-tailed, imbalanced medical data[26], our integrated patient-level stratified class-balancing pipeline stabilizes feature mapping. This allows our optimized architecture to sustain a minimal 2.1% mAP drop, outperforming standard unstratified compression methods.

5) Clinical Implications, Limitations, and Future Work: Compressing the model to a sub-60MB footprint with a 6.24 ms latency enables offline clinical screening on resource-constrained localized computing hardware common in rural, low-connectivity facilities. However, a key limitation is the lack of physical profiling on specialized low-power edge hardware; future work must transition from CPU simulations to direct edge deployment to map physical thermal, power, and energy-efficiency metrics. Additionally, the model must be validated on multi-demographic datasets to prevent domain shifts and extended to multi-class detection (e.g., Tuberculosis and COVID-19) to expand its clinical utility.

IV. CONCLUSION

This study demonstrates the practical feasibility of deploying advanced deep learning models to help address the global health burden of pneumonia, particularly in settings where access to radiologists is limited. By refining the YOLOv5lu architecture through a combined strategy of structured pruning and INT8 quantization, we achieved nearly a twofold reduction in model size and accelerated inference to approximately 6.24 ms per image on standard hardware. Importantly, these efficiency gains were obtained with minimal loss of performance compared to the baseline, indicating that lightweight optimization itself is not the primary constraint on diagnostic accuracy. Research on YOLO model compression demonstrates that pruning and quantization can significantly reduce model complexity while maintaining performance. Advanced compression frameworks can achieve 73.51% parameter reduction while reducing detection accuracy by 2.7%[25]. This validates our approach, the low accuracy we observed is fundamentally not a consequence of the compression techniques themselves, but rather reflects underlying challenges in the dataset and task domain. A significant challenge we faced was that the models, both before and after optimization, showed low accuracy scores (measured by Mean Average Precision, or mAP). This problem comes from the nature of the

RSNA dataset itself. In many pneumonia cases, the signs in the lungs appear as hazy or blurry patches with unclear edges, and they often blend into the surrounding tissue. That makes them much harder to detect than the sharp, well defined objects usually found in standard computer vision tasks. Future work should therefore focus on strengthening feature representation, for example through attention mechanisms or higher resolution inputs, to raise the performance ceiling in this domain. At the same time, expanding the framework to multi class detection tasks such as Tuberculosis and COVID-19 will be essential to broaden its clinical relevance.

REFERENCE

[1] C. King *et al.*, "Paediatric pneumonia research priorities in the context of COVID-19: An eDelphi study," *J. Glob. Health*, vol. 12, p. 09001, 2022.
[2] World Health Organization, "Stakeholder Consultative Meeting on Prevention and Management of Childhood Pneumonia and Diarrhoea Report, 12–14 October 2021." WHO, Geneva, Switzerland, 2022.
[3] C. Cairns and K. Kang, "National Hospital Ambulatory Medical Care Survey: 2022 emergency department summary tables. Tables 11." 2022. [Online]. Available: https://www.cdc.gov/nchs/data/nhamcs/web_tables/2022-nhamcs-ed-web-tables.pdf
[4] Centers for Disease Control and Prevention, National Center for Health Statistics, "National Vital Statistics System, Mortality 2018–2023 on CDC WONDER Online Database." 2024. [Online]. Available: http://wonder.cdc.gov/ucd-icd10-expanded.html
[5] A. H. Attaway, R. G. Scheraga, A. Bhimraj, M. Biehl, and U. Hatipoğlu, "Severe covid-19 pneumonia: pathogenesis and clinical management," *BMJ*, p. n436, Mar. 2021, doi: 10.1136/bmj.n436.
[6] A. J. Amron, C. Paramita, and P. Šolić, "Interpretable Hybrid YOLOv8s-GWO Framework for Bounding-Box Viral Pneumonia Detection on Kaggle Chest X-ray Images," vol. 6, no. 6, 2025.
[7] S. Singh, M. Kumar, A. Kumar, B. K. Verma, K. Abhishek, and S. Selvarajan, "Efficient pneumonia detection using Vision Transformers on chest X-rays," *Sci. Rep.*, vol. 14, no. 1, p. 2487, Jan. 2024, doi: 10.1038/s41598-024-52703-2.
[8] T. O. Togunwa, A. O. Babatunde, O. E. Fatade, R. Olatunji, G. Ogbole, and A. Falade, "Detection of pneumonia in children through chest radiographs using artificial intelligence in a low-resource setting: A pilot study," *PLOS Digit. Health*, vol. 4, no. 9, p. e0000713, Sep. 2025, doi: 10.1371/journal.pdig.0000713.
[9] A. Stein *et al.*, "RSNA Pneumonia Detection Challenge." Kaggle, 2018. [Online]. Available: https://kaggle.com/competitions/rsna-pneumonia-detection-challenge
[10] M. Larobina, "Thirty Years of the DICOM Standard," *Tomography*, vol. 9, no. 5, pp. 1829–1838, Oct. 2023, doi: 10.3390/tomography9050145.
[11] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *Proc. IEEE*, vol. 111, no. 3, pp. 257–276, Mar. 2023.
[12] M. Nawaz, R. Qureshi, M. A. Teevno, and A. R. Shahid, "Object detection and segmentation by composition of fast fuzzy C-mean clustering based maps," *J. Ambient Intell. Humaniz. Comput.*, vol. 14, no. 6, pp. 7173–7188, Jun. 2023.
[13] W. T. D. Tjahjono, C. Paramita, C. Supriyanto, A. J. Savicevic, S. Rakasiwi, and A. A. Haresta, "YOLOv10x Model for Accurate First Detection of Skin Diseases from Dermoscopic Objects," in *2025 International Conference on Smart Computing, IoT and Machine Learning (SIML)*, Surakarta, Indonesia: IEEE, Jun. 2025, pp. 1–6. doi: 10.1109/SIML65326.2025.11080864.
[14] M. A. Al-antari, S.-M. Han, and T.-S. Kim, "Evaluation of deep learning detection and classification towards computer-aided diagnosis of breast lesions in digital X-ray mammograms," *Comput. Methods Programs Biomed.*, vol. 196, p. 105584, Nov. 2020, doi: 10.1016/j.cmpb.2020.105584.
[15] Y. E. Almalki *et al.*, "Deep Learning Models for Classification of Dental Diseases Using Orthopantomography X-ray OPG Images," *Sensors*, vol. 22, no. 19, p. 7370, Sep. 2022, doi: 10.3390/s22197370.
[16] A. Baccouche, B. Garcia-Zapirain, Y. Zheng, and A. S. Elmaghraby, "Early detection and classification of abnormality in prior mammograms using image-to-image translation and YOLO techniques," *Comput. Methods Programs Biomed.*, vol. 221, p. 106884, Jun. 2022, doi: 10.1016/j.cmpb.2022.106884.
[17] M. Dobrovolny, J. Benes, J. Langer, O. Krejcar, and A. Selamat, "Study on Sperm-Cell Detection Using YOLOv5 Architecture with Labaled Dataset," *Genes*, vol. 14, no. 2, p. 451, Feb. 2023, doi: 10.3390/genes14020451.
[18] Z. Han, H. Huang, Q. Fan, Y. Li, Y. Li, and X. Chen, "SMD-YOLO: An efficient and lightweight detection method for mask wearing status during the COVID-19 pandemic," *Comput. Methods Programs Biomed.*, vol. 221, p. 106888, Jun. 2022, doi: 10.1016/j.cmpb.2022.106888.

[19] Cinantya Paramita, Catur Supriyanto, Amalia, and Khalivio Rahmyanto Putra, "Comparative Analysis of YOLOv5 and YOLOv8 Cigarette Detection in Social Media Content," *Sci. J. Inform.*, vol. 11, no. 2, pp. 341–352, May 2024, doi: 10.15294/sji.v11i2.2808.
[20] T. Diwan, G. Anirudh, and J. V. Tembhurne, "Object detection using YOLO: Challenges, architectural successors, datasets and applications," *Multimed. Tools Appl.*, vol. 82, no. 6, pp. 9243–9275, Mar. 2023.
[21] W. Fang, L. Wang, and P. Ren, "Tinier-YOLO: A real-time object detection method for constrained environments," *IEEE Access*, vol. 8, pp. 1935–1944, 2020.
[22] C. Paramita, C. Supriyanto, P. Šolić, C. Wada, and A. A. Dzaky, "Performance Evaluation of YOLOv8 Models for Multi-Class Skin Lesion Detection from Dermoscopic Images," in *2025 International Conference on Smart Computing, IoT and Machine Learning (SIML)*, Surakarta, Indonesia: IEEE, Jun. 2025, pp. 1–6. doi: 10.1109/SIML65326.2025.11080819.
[23] T. Liang, J. Glossner, L. Wang, S. Shi, and X. Zhang, "Pruning and quantization for deep neural network acceleration: A survey," *Neurocomputing*, vol. 461, pp. 370–403, Oct. 2021, doi: 10.1016/j.neucom.2021.07.045.
[24] A. Kuzmin, M. Nagel, M. van Baalen, A. Behboodi, and T. Blankevoort, "Pruning vs Quantization: Which is Better?," 2023, *arXiv*. doi: 10.48550/ARXIV.2307.02973.
[25] M. Sabaghian, M. A. Keyvanrad, and S. M. Moghadami, "A Novel Compression Framework for YOLOv8: Achieving Real-Time Aerial Object Detection on Edge Devices via Structured Pruning and Channel-Wise Distillation," 2025, *arXiv*. doi: 10.48550/ARXIV.2509.12918.
[26] G. Holste *et al.*, "How Does Pruning Impact Long-Tailed Multi-Label Medical Image Classifiers?," vol. 14224, 2023, pp. 663–673. doi: 10.1007/978-3-031-43904-9_64.