Automatic Pill Detection Using Faster R-CNN with an AlexNet Backbone

Dinial Utami Nurul Qomariah[1], Ade Irma Elvira[2], Arvita Agus Kurniasar[3], Bima Wahyu Maulana[4]

[1] Department of Information Technology, Politeknik Negeri Jember, Jember, Indonesia,
[2,3,4] Department of Biology, Universitas Indonesia, Depok, Indonesia
Email: dinial.utami@polije.ac.id

ABSTRAK

Deteksi objek merupakan komponen penting dalam pengembangan sistem otomatis di bidang kesehatan, khususnya pada sektor farmasi, seperti proses identifikasi dan pengelolaan obat. Tantangan utama dalam sistem deteksi obat berbasis citra adalah mencapai tingkat akurasi yang tinggi serta kemampuan generalisasi yang baik terhadap variasi bentuk, warna, dan kondisi pencahayaan obat. Penelitian ini menerapkan metode Faster R-CNN dengan backbone AlexNet untuk mendeteksi dan mengklasifikasikan objek obat pada citra digital. Proses pelatihan dilakukan dengan beberapa skenario jumlah epoch untuk mengevaluasi pengaruh durasi pelatihan terhadap kinerja model. Hasil evaluasi menunjukkan bahwa model mampu mencapai tingkat akurasi hingga 98%, yang menandakan kemampuan deteksi obat yang sangat baik. Peningkatan jumlah pelatihan memberikan performa yang lebih stabil dan konsisten dalam mengenali objek obat. Berdasarkan hasil eksperimen, dapat disimpulkan bahwa Faster R-CNN berbasis AlexNet efektif diterapkan pada aplikasi farmasi, terutama untuk mendukung sistem distribusi, pengemasan, dan penghitungan obat yang membutuhkan ketelitian dan keandalan tinggi.

Kata Kunci: *Deteksi Objek, Deteksi Obat, Faster R-CNN, AlexNet, Deep Learning.*

ABSTRACT

Object detection is a crucial component in the development of automated systems in the healthcare domain, particularly in pharmaceutical applications such as pill identification and management. One of the main challenges in image-based pill detection systems is achieving high accuracy and robust generalization under variations in pill shape, color, and illumination conditions. This study applies the Faster R-CNN framework with an AlexNet backbone to detect and classify pill objects in digital images. The model is trained using multiple epoch configurations to analyze the effect of training duration on detection performance. Experimental results show that the proposed approach achieves an accuracy of up to 98%, demonstrating strong detection capability. Increasing the number of training epochs improves the stability and consistency of pill recognition. These results indicate that AlexNet-based Faster R-CNN is effective for pharmaceutical applications, particularly in drug distribution, packaging, and pill counting systems that require high precision and reliability.

Keywords: *Object Detection, Pill Detection, Faster R-CNN, AlexNet, Deep Learning.*

INTRODUCTION

Errors in pill counting and drug distribution remain significant issues that affect patient safety and healthcare service efficiency, potentially leading to dosage errors, treatment delays, and economic losses (Bates et al., 1997; Morimoto et al., 2004). Automation of pill counting processes in pharmacies and pharmaceutical production lines aims to reduce manual errors by improving consistency (Heo et al., 2023) and processing speed (S.MANIKANDAN et al., 2025). In the field of computer vision, two-stage object detection approaches (Babasaheb, 2023; Chughtai et al., 2019), such as Faster R-CNN (Ren et al., 2017) with an AlexNet backbone (Krizhevsky et al., 2012), provide accurate detection by separating region proposal and classification stages, making them suitable for pharmaceutical applications that require high precision in pill identification and counting (Tan et al., 2021).

Technical challenges in pill counting tasks include the small size of pill objects (Nikouei et al., 2025), inter-pill occlusion, variations in illumination and background, and reflective surfaces that may cause false positive and false negative detections (Akyon et al., 2022). Previous studies on small object detection and object counting propose several effective mitigation strategies (Qomariah et al., 2021; Yang et al., 2023), including multi-scale feature representation, increased input image resolution, loss functions that emphasize small objects, and domain-specific data augmentation to improve robustness under real-world conditions (Li et al., 2022; Liang et al., 2022).

Research in pharmaceutical and healthcare domains shows that the integration of deep learning techniques, particularly object detection combined with pill identification, achieves practical accuracy and reduces the risk of medication error (Al-Hussaeni et al., 2023; Chiu, 2024). Comparative studies of object detection models such as RetinaNet, SSD, and Faster R-CNN report that two-stage detectors provide high localization accuracy, making them suitable for applications that require precise object recognition or segmentation (Abdullah et al., 2018; Qomariah et al., 2025). Classical image processing approaches remain applicable in specific cases, such as pill detection in blister packaging with relatively uniform shapes; however, these methods generally exhibit limited generalization to diverse real-world scenarios (M et al., 2023).

For pill counting, the literature typically distinguishes two main approaches: (a) count-by-detection, which estimates the number of pills by counting detected bounding boxes, and (b) regression- or density-based methods that directly predict object counts from images. Hybrid strategies that combine object detection for localization with post-processing counting or customized non-maximum suppression demonstrate strong performance in dense and overlapping pill scenarios (Cohen et al., 2017; Liu et al., 2018). Deployment on edge devices requires lightweight models and optimization techniques, such as pruning, quantization, and ONNX-based export, to ensure efficient inference while maintaining detection accuracy (Guerrouj et al., 2025).

Considering the existing research gap, particularly the limited evaluation under dense multi-pill conditions and the lack of comprehensive counting-specific metrics, this study proposes the design, implementation, and evaluation of an Automatic Pill Counting system based on Faster R-CNN with an AlexNet backbone (Krizhevsky et al., 2012; Ren et al., 2017). The proposed system focuses on realistic multi-pill dataset construction and augmentation, optimization of Faster R-CNN for small pill detection, comprehensive evaluation of detection and counting performance, and deployment-oriented optimization to support reliable integration into pharmacy management systems.

Material (Dataset)

The PillBox (retired) dataset is a reference collection of solid oral medication images developed by the National Library of Medicine (NLM) (Pillbox, 2025). The images are acquired using a professional studio imaging system with high-resolution digital cameras, controlled illumination, homogeneous backgrounds, and consistent object orientations, resulting in high-quality images with resolutions of up to 1600×1600 pixels and full-color representation. The dataset is accompanied by comprehensive metadata, including physical dimensions, geometric shape, dominant color, imprint text, and linkage to drug ontology through RxNorm.

In this study, the PillBox dataset is used as a baseline reference dataset for developing an automatic pill detection and counting system based on Faster R-CNN with an AlexNet backbone. Since most images contain a single object per frame, data augmentation is applied to simulate real-world conditions, including Flippin in H-flip and V-flip, rotation, brightness and zooming. The high resolution of the original images enables controlled downsampling to match the model input size while preserving critical visual features. Consequently, the PillBox dataset serves as a reliable ground-truth reference for model training and evaluation prior to deployment on more complex, non-controlled images.

RESEARCH METHODS

Faster R-CNN is a two-stage object detection framework that integrates feature extraction, region proposal generation, and object classification with bounding box regression in an end-to-end manner (Ren et al., 2017). Given an input image I, a convolutional neural network backbone is employed to extract high-level visual representations, which can be expressed in equation 1.

$$F = B(I) \qquad (1)$$

where $B(.)$ denotes the backbone network and F represents the resulting feature map used for subsequent detection stages.

In this study, AlexNet is adopted as the backbone network of Faster R-CNN to extract discriminative features from pill images. AlexNet consists of five convolutional layers followed by three fully connected layers, utilizing large convolution kernels in the early layers to capture global visual patterns and smaller kernels in deeper layers to encode more detailed semantic features. Each convolutional layer is followed by a ReLU activation function, while max-pooling layers are applied to reduce spatial resolution and improve translation invariance. Dropout and local response normalization are employed to enhance training stability and reduce overfitting (Krizhevsky et al., 2012).

Based on the extracted feature map F, the Region Proposal Network (RPN) generates a set of candidate object regions by predicting objectness scores and bounding box offsets for predefined anchor boxes. The RPN is optimized using a multi-task loss function that combines classification and regression terms, formulated in equation 2.

$$\mathcal{L}_{RPN} = \mathcal{L}_{cls}(p, p^*) + \leftthreetimes \mathcal{L}_{reg}(t, t^*) \ (2)$$

where p and p^* represent the predicted and ground-truth objectness labels, and t and t^* denote the predicted and reference bounding box parameters.

The generated region proposals are then aligned to a fixed-size feature representation using RoI Align and forwarded to the Fast R-CNN detection head for final object classification and bounding box refinement. The overall optimization objective of Faster R-CNN is defined in equation 3.

$$\mathcal{L} = \mathcal{L}_{RPN} + \mathcal{L}_{det} \ (3)$$

This two-stage detection strategy enables precise localization and classification of small objects, such as pharmaceutical pills, making Faster R-CNN with an AlexNet backbone well suited for automated pill detection and counting systems that require high accuracy and reliability.

A. Evaluation Metrics

The performance of the proposed Faster R-CNN with AlexNet backbone (Krizhevsky et al., 2012; Ren et al., 2017) is evaluated using accuracy, precision, and recall metrics to assess its detection reliability in pharmaceutical pill identification. Accuracy measures the overall correctness of the model by comparing the number of correct predictions to the total number of evaluated samples, and it is defined in equation 4.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

where TP denotes true positives, TN true negatives, FP false positives, and FN false negatives. This metric provides a general indication of model effectiveness in correctly detecting and rejecting pill objects.

Precision and recall are employed to evaluate the model's capability in identifying pill objects accurately and completely. Precision quantifies the proportion of correctly detected pill instances among all detected instances, while recall measures the proportion of correctly detected pill instances relative to all ground-truth instances. These metrics are formulated in equation 5 and 6.

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

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

High precision indicates a low false positive rate, whereas high recall reflects the model's robustness in minimizing missed detections. Together, these metrics provide a comprehensive evaluation of the Faster R-CNN model performance in pharmaceutical applications, where accurate and reliable pill detection is critical for reducing medication errors.

RESULTS AND DISCUSSION

The dataset used in this study is derived from the PillBox (Retired) collection, which was previously developed and maintained by the U.S. National Library of Medicine (NLM). This dataset contains thousands of standardized images of prescription pills with diverse visual characteristics, including variations in shape, color, size, and imprint patterns used for drug identification. Each image is provided at a consistent resolution and accompanied by descriptive labels prior to the dataset being retired. The high visual diversity of the PillBox dataset makes it well suited for training deep learning–based object detection models, particularly Faster R-CNN with an AlexNet backbone (Krizhevsky et al., 2012; Ren et al., 2017), to improve generalization performance in pharmaceutical pill detection tasks.

In the experimental setup, the dataset is randomly partitioned into training, validation, and testing subsets with ratios of 70%, 20%, and 10%, respectively. Each image is consistently paired with its corresponding annotation file to preserve label integrity across all splits, and the dataset is shuffled prior to partitioning to mitigate sampling bias. To enhance data diversity and improve model generalization, several data augmentation techniques are applied to the training set, including horizontal and vertical flipping, small-angle rotation, brightness adjustment, and slight zooming, as illustrated in Figure 1. The augmented dataset is then organized into separate directory structures for images and labels, enabling systematic model training, hyperparameter tuning, and objective evaluation of the Faster R-CNN model on unseen pill images.

Faster R-CNN is implemented using AlexNet as the feature extraction backbone. The pretrained AlexNet model is initialized with ImageNet weights, and only the convolutional feature extractor is utilized to generate feature maps, with the output channel size set to 256 to match the detection head requirements. A custom Region Proposal Network (RPN) is defined using an anchor generator with five anchor scales (32, 64, 128, 256, 512) and th sree aspect ratios (0.5, 1.0, 2.0) to accommodate pills of varying sizes and shapes. For region feature extraction, a MultiScale RoI Align layer with an output size of 7×7 is applied to the single feature map produced by AlexNet. The complete Faster R-CNN model integrates the backbone, RPN, and RoI pooling components, enabling end-to-end training for pill detection and classification. This configuration balances detection accuracy and computational efficiency, making it suitable for pharmaceutical image analysis.

the Faster R-CNN model with an AlexNet backbone is trained using a stochastic gradient descent (SGD) optimizer with a learning rate of 0.005, momentum of 0.9, and weight decay of 0.0005, which provides stable convergence during training. The model is trained for 10 epochs on a GPU when available, and a StepLR scheduler is applied to reduce the learning rate by a factor of 0.1 every five epochs, enabling finer parameter updates in later training stages. During each iteration, the total loss is computed as the sum of classification, bounding box regression, and RPN losses, and backpropagation is performed end-to-end. The reported training loss consistently decreases across epochs, indicating effective learning and convergence of the Faster R-CNN model. This training strategy contributes to the reliable detection performance observed in the evaluation metrics, including accuracy, precision, and recall, on the test dataset as shown in Table 1.

Table 1. Performance Evaluation Comparison of the Proposed Faster R-CNN with Alexnet

Method.	Epoch	Percentage			Minute
		Precision	Recall	Accuracy	Time Computation
FasterRCNN+Resnet50	10	95	92	93.5	233
FasterRCNN+Resnet50	50	96	96	96	978
FasterRCNN+Alexnet	10	**98**	96	97	**67**
FasterRCNN+Alexnet	50	**98**	**99.7**	**98.8**	160

Table 1 presents a performance evaluation comparison between Faster R-CNN models using ResNet50 and AlexNet backbones under different training epochs. The results indicate that Faster R-CNN with an AlexNet backbone consistently outperforms the ResNet50-based model in terms of precision, recall, and overall accuracy. At 10 epochs, Faster R-CNN + AlexNet achieves a precision of 98% and a recall of 96%, resulting in an accuracy of 97%, while also requiring significantly less computation time (67 minutes) compared to ResNet50 (233 minutes). When trained for 50 epochs, the AlexNet-based model further improves recall to 99.7% and reaches the highest accuracy of 98.8%, demonstrating superior detection performance with relatively low computational cost.

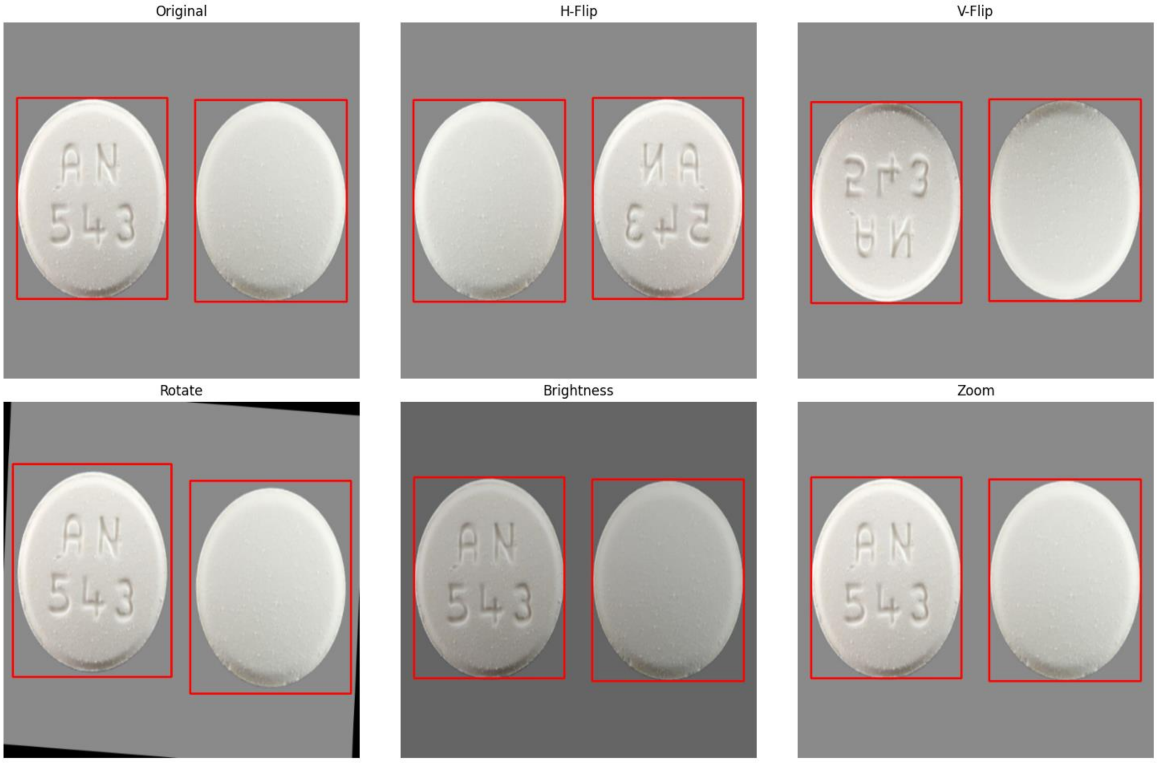

Figure 1. Augmentation Results.

Discussion

The comparative results in Table 1 highlight the effectiveness of using AlexNet as a backbone for Faster R-CNN in pill detection tasks (Krizhevsky et al., 2012; Ren et al., 2017). Despite ResNet50 (He et al., 2016) being a deeper network, its performance gains are marginal compared to the substantial increase in computational time. In contrast, AlexNet provides a favorable trade-off between accuracy and efficiency, achieving higher precision and recall while significantly reducing training time. The improvement observed at 50 epochs suggests that the AlexNet-based model benefits from longer training without overfitting, leading to better generalization. These findings indicate that Faster R-CNN with AlexNet is more suitable for practical pharmaceutical applications, where high detection accuracy and computational efficiency are critical for real-world deployment.

The detection results shown in Figure 2. indicate that the proposed Faster R-CNN model with an AlexNet backbone is capable of accurately localizing and classifying pill objects in a single image. The model successfully identifies two pill instances as Class 1 with a confidence score of 1.00, demonstrating high detection reliability. The bounding boxes closely follow the pill boundaries, even in the presence of background elements such as measurement scales, indicating that the region proposal and feature extraction processes are robust to contextual noise. Moreover, the correct detection of pills with different surface imprints confirms that the model effectively learns discriminative features related to pill shape and texture rather than relying solely on imprint patterns. These qualitative results are consistent with the quantitative evaluation and further validate the effectiveness of the proposed approach for practical pill detection applications.

CONCLUSION

This research confirms that the Faster R-CNN model with an AlexNet backbone is an effective and efficient solution for pharmaceutical pill detection using the PillBox (Retired) dataset. The experimental results demonstrate that the proposed approach achieves high detection performance, with precision, recall, and accuracy reaching up to 98%, 99.7%, and 98.8%, respectively. Compared to the ResNet50-based Faster R-CNN, the AlexNet backbone provides superior performance while requiring substantially less computation time, making it more suitable for practical applications. The incorporation of data augmentation techniques further improves model generalization by exposing the network to variations in orientation, illumination, and scale, which are commonly encountered in real-world pill images.

Furthermore, qualitative detection results show that the proposed model can accurately localize and classify pill objects with high confidence, even in the presence of background noise such as measurement scales and varying imprint patterns. The bounding boxes closely align with pill boundaries, indicating robust region proposal and feature extraction capabilities. Overall, the balance between high accuracy and computational efficiency achieved by Faster R-CNN with AlexNet highlights its potential for deployment in real-world pharmaceutical systems, such as automated pill identification and verification, where reliability and efficiency are critical requirements.

REFERENCES

Abdullah, M., Qomariah, D., & Farosanti, L. (2018). AUTOMATIC DETERMINATION OF SEEDS FOR RANDOM WALKER BY SEEDED WATERSHED TRANSFORM FOR TUNA IMAGE SEGMENTATION. *Jurnal Ilmu Komputer Dan Informasi*, *11*, 52. https://doi.org/10.21609/jiki.v11i1.468

Akyon, F. C., Onur Altinuc, S., & Temizel, A. (2022). Slicing Aided Hyper Inference and Fine-Tuning for Small Object Detection. *2022 IEEE International Conference on Image Processing (ICIP)*, 966–970. https://doi.org/10.1109/ICIP46576.2022.9897990

Al-Hussaeni, K., Karamitsos, I., Adewumi, E., & Amawi, R. M. (2023). CNN-Based Pill Image Recognition for Retrieval Systems. *Applied Sciences*, *13*(8). https://doi.org/10.3390/app13085050

Babasaheb, K. D. (2023). *A deep learning based drug pill recognition system*. *7*(6).

Bates, D. W., Spell, N., Cullen, D. J., Burdick, E., Laird, N., Petersen, L. A., Small, S. D., Sweitzer, B. J., & Leape, L. L. (1997). The costs of adverse drug events in hospitalized patients. Adverse Drug Events Prevention Study Group. *JAMA*, *277*(4), 307–311.

Chiu, Y.-J. (2024). Automated medication verification system (AMVS): System based on edge detection and CNN classification drug on embedded systems. *Heliyon*, *10*(9), e30486. https://doi.org/10.1016/j.heliyon.2024.e30486

Chughtai, M., Raja, G., Mir, J., & Shaukat, F. (2019). *An Efficient Scheme for Automatic Pill Recognition Using Neural Networks*. *56*, 42–48.

Cohen, J. P., Boucher, G., Glastonbury, C. A., Lo, H. Z., & Bengio, Y. (2017). Count-ception: Counting by Fully Convolutional Redundant Counting. *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, 18–26. https://doi.org/10.1109/ICCVW.2017.9

Guerrouj, F. Z., Rodriguez Florez, S., El Ouardi, A., Abouzahir, M., & Ramzi, M. (2025). Quantized Object Detection for Real-Time Inference on Embedded GPU Architectures. *International Journal of Advanced Computer Science and Applications*, *16*. https://doi.org/10.14569/IJACSA.2025.0160503

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Heo, J., Kang, Y., Lee, S., Jeong, D.-H., & Kim, K.-M. (2023). An Accurate Deep Learning-Based System for Automatic Pill Identification: Model Development and Validation. *Journal of Medical Internet Research*, *25*, e41043. https://doi.org/10.2196/41043

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems 25 (NIPS 2012)*, 1–9.

Li, L., Li, B., & Zhou, H. (2022). Lightweight multi-scale network for small object detection. *PeerJ. Computer Science*, *8*, e1145. https://doi.org/10.7717/peerj-cs.1145

Liang, Y., Han, Y., & Jiang, F. (2022). *Deep Learning-based Small Object Detection: A Survey*. 432–438. https://doi.org/10.1145/3532213.3532278

Liu, J., Gao, C., Meng, D., & Hauptmann, A. G. (2018). DecideNet: Counting Varying Density Crowds Through Attention Guided Detection and Density Estimation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5197–5206. https://doi.org/10.1109/CVPR.2018.00545

M, S., G, E. R., D, A., Akshaya, S., Uthaman, A. C., & Sridhar, S. (2023). Detection and Identification of Pills using Machine Learning Models. *2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN)*, 1–6. https://doi.org/10.1109/ViTECoN58111.2023.10157873

Morimoto, T., Gandhi, T., Seger, A. C., Hsieh, T., & Bates, D. (2004). Adverse drug events and medication errors: Detection and classification methods. *Quality & Safety in Health Care*, *13*, 306–314. https://doi.org/10.1136/qhc.13.4.306

Nikouei, M., Baroutian, B., Nabavi, S., Taraghi, F., Aghaei, A., Sajedi, A., & Ebrahimi Moghaddam, M. (2025). *Small Object Detection: A Comprehensive Survey on Challenges, Techniques and Real-World Applications*. https://doi.org/10.48550/arXiv.2503.20516

Pillbox. (2025). *Pillbox - Archived Data*. Open Data Portal.

Qomariah, D. U. N., Tjandrasa, H., & Alam, B. R. (2021). Hemorrhage Segmentation in Retinal Images Using Modified FCN-8. *2021 Fourth International Conference on Vocational Education and Electrical Engineering (ICVEE)*, 1–6. https://doi.org/10.1109/ICVEE54186.2021.9649686

Qomariah, D. U. N., Tjandrasa, H., & Elvira, A. I. (2025). Retinal Blood Vessel Segmentation Based on Encoder and Decoder Networks Using Weighted Cross Entropy Loss Function. *ADALAH*, *9*(6).

Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(6), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031

S.MANIKANDAN, M., KUMAR, A., KUMAR, V., NITHISH, V., & KRISHNA, G. (2025). AUTOMATED PILL RECOGNITION AND CLASSIFICATION USING CONVOLUTIONAL NEURAL NETWORKS. *International Journal of Engineering Research and Science & Technology*, *21*, 654–660. https://doi.org/10.62643/ijerst.2025.v21.n4.pp654-660

Tan, L., Huangfu, T., Wu, L., & Chen, W. (2021). Comparison of RetinaNet, SSD, and YOLO v3 for real-time pill identification. *BMC Medical Informatics and Decision Making*, *21*(1), 324. https://doi.org/10.1186/s12911-021-01691-8

Yang, X., Yan, J., Liao, W., Yang, X., Tang, J., & He, T. (2023). SCRDet++: Detecting Small, Cluttered and Rotated Objects via Instance-Level Feature Denoising and Rotation Loss Smoothing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(2), 2384–2399. https://doi.org/10.1109/TPAMI.2022.3166956