Proceedings of the 4th International Conference on Applied Artificial Intelligence and Computing (ICAAIC-2025)
DVD Part Number: CFP25BC3-DVD; ISBN: 979-8-3315-6586-2

AI-Driven Object Detection Using YOLOv7 and Enhanced Small-Object Recognition with Vision Transformers for Real-Time Surveillance Solutions

Keerthana.R [1],
Department of Computer Science and Engineering,
K.S.R. College of Engineering Tiruchengode
keerthiksrcecse@gmail.com

Dr.E.Baby Anitha [2],
Department of Computer Science and Engineering,
K.S.R College of Engineering Tiruchengode.
ebabyanitha@gmail.com

Karthika.J [3],
Department of Computer Science and Engineering,
*Mahendra Engineering College for Women,*Tiruchengode.
j.karthikasumathi@gmail.com

Agalya.M [4]
Department of Computer Science and Engineering
*Vivekanandha College of Technology for Women,*Tiruchengode,
agalkalai@gmail.com

Haritha.T [5]
Department of Computer Science and Technology,
Vivekanandha College of Engineering for Women,, Tiruchengode.
haritha@vcew.ac.in

R.V.Viswanathan [6],
Department of IT,
Kongunadu college of Engineering & Technology, Thottiyam,
Trichy,
viswamecse@gmail.com

Abstract—Aim: The aim of this study is to create an object identification framework that enhances current models by utilizing YOLOv4, YOLOv6, YOLOv7, and Vision Transformers. The main objectives are to improve computational efficiency, small object recognition, and real-time detection capabilities. Materials and Methods: The experimental setup consists of two groups: Group 1 (Control) utilizes YOLOv4 and YOLOv6 on the PASCAL VOC dataset, comprising 16,551 training and validation images. However, it exhibits limitations in accuracy and scalability and Group 2 (Intervention), refers to 200,000 Images and 80 object categories, which introduces a novel technique that enhances small and occluded object detection with a 95% confidence interval through the integration of Vision Transformers and hybrid architectures utilizing YOLOv7. Results: The proposed YOLOv7 + Vision Transformers framework achieves higher accuracy than YOLOv7. Its precision ranges from 94.8% to 95.0% with accuracy 94.5%, outperforming YOLOv7's 91.2% to 93.1%.The optimal inference time is 20.8 ms, significantly lower than YOLOv7's 21.1 ms to 23.9 ms,with a significance of 0.05.Conclusion: In this work, its is observed that the YOLOv7 and Vision Transformer has significantly provide a better accuracy and establishing a robust solution for enhanced detection capabilities.

Keywords—Object Detection, Machine Learning, Deep Learning, YOLO7, Vision Transformer, Data Augmentation ,Mean Average Precision (mAP), Real-Time Processing

I. INTRODUCTION

Finding and locating objects in pictures or videos is the main goal of object detection, a fundamental problem in computer vision. (Guo et al. 2025). It plays a critical role in enabling machines to interpret and interact with their surroundings. Applications of object detection span across surveillance, autonomous vehicles, robotics, and inventory management. The evolution of this field, from traditional handcrafted feature-based techniques to deep learning-based methods, underscores its importance in modern computer vision systems. The advent of deep neural network architectures like Vision Transformers (ViTs) and the YOLO (You Only Look Once) family, which includes YOLOv4, YOLOv6, and YOLOv7, has fueled recent advances in object detection. (StoryBuddiesPlay 2025). These models have set benchmarks for their speed, accuracy, and ability to detect objects of varying sizes and handle challenging scenarios like occlusion and cluttered backgrounds. For example, YOLOv7 has demonstrated an average precision (AP) improvement of +3.5% over YOLOv4 while maintaining real-time processing capabilities(Lim et al. 2025). Techniques like data augmentation and ensemble learning have further enhanced the robustness of these systems, enabling them to generalize better to diverse environments (Ikram and Imran 2025). Despite these advances, challenges persist in detecting small and occluded objects, optimizing models for computational efficiency, and achieving scalability across hardware platforms (De Geoffroy and Wignall 2012).

.

II. RELATED WORKS

The total number of articles published on this topic over the last five years is more than 4457 papers in IEEE Xplore, 901 papers in Google Scholar, and 105 papers in academia.edu. Recent advancements in object detection have been driven by remarkable progress in deep learning, particularly through the development of the YOLO family of models (YOLOv4, YOLOv6, and YOLOv7) and the adaptation of Vision Transformers. These innovations have effectively addressed critical challenges such as detecting small objects , managing occlusions, and meeting the demands of real-time processing (Guo et al. (2025). Pioneering contributions from researchers Q. Zhang et al. (2024). Introduced significant improvements in areas like unsupervised learning , data augmentation, and optimized deployment techniques are concentrated (Raj and Gayathri 2024). Object detection, a key area in computer vision, has made substantial progress with advancements in deep learning models like YOLO and Vision Transformers have data augmentation which achieved exceptional real-time performance, accurately identifying and classifying objects in images across various applications Sharma & Rajesh Kumar (2010). However, challenges remain while Detecting small objects, which is difficult due to insufficient pixel information, while occlusions can impair the accurate localization of

partially hidden objects Schoeman et al.(2024) Additionally, the computational intensity of many models limits their use on devices with constrained resources which are mentioned in the paper (S. Tang et al. 2025). Additionally, integrating object detection with tasks like image segmentation and scene understanding holds great promise for expanding its applications and impact as mentioned (Gu et al. 2025).

From previous findings, YOLO models and Vision Transformers have achieved addressing challenges like real-time processing and small object detection. However, issues with crowded environments and computational limits remain. Future research aims to enhance robustness, efficiency, and integration with image segmentation.

The goal of this research is to develop a computer vision-based system that combines worker identification and helmet wearing detection to promote safety in industrial and construction settings. The current approach emphasizes precision, recall, and performance under controlled conditions by using CNN and sliding window techniques on a dataset of 5000 pictures. The suggested approach focuses on real-time performance and limitations such occlusions, light conditions, and flexibility by combining YOLOv11 with Grassmann algorithms. Direct comparison was made possible by the standard dataset, demonstrating the improved performance and capacity of the suggested approach[23].

III. Materials and Methods

The study for object detection research includes deep learning frameworks , large amounts of data storage, and high-performance computing (GPUs). Experiments include comparing various architectures, assessing performance using metrics like mAP and IoU, and training models on datasets (Sharma, Sengar, and Singh 2023). Deep learning frameworks like TensorFlow/PyTorch and common assessment metrics (like mAP) are required for object identification research with deep learning. The model makes use of two primary object identification datasets: PASCAL VOC (Visual Object Classes), which is used to assess particular item categories, and COCO (Common Objects in Context), which is renowned for its vast and varied sceneries with data augmentation.

Group 1: The study's foundational benchmark for model evaluation in the exsisitng system is the PASCAL VOC dataset, notably the combined training and validation split consisting of 16,551 images. With an inference time of roughly 24.5 milliseconds per image, two models, YOLOv4 and YOLOv6, are trained on this dataset, suggesting a throughput appropriate for near-real-time applications. YOLOv6 performs marginally better than the other. This is due to small architectural and training-strategy enhancements that increase detection speed or accuracy without significant trade-offs. The choice to base "Group 1's" sample size on these earlier benchmarks guarantees that the experimental design is in line with accepted literature, allowing for a meaningful comparison of outcomes.

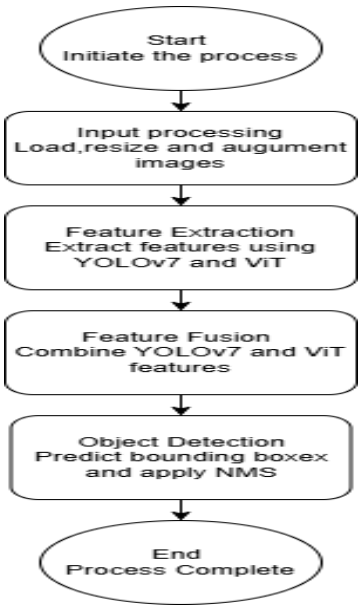
This research lowers variability caused by dataset scale or architecture novelty by anchoring the dataset size, model selections, and inference metrics to known values, concentrating instead on the effects of methodological modifications. Furthermore, reliability and comparability between investigations are supported by the use of a standard benchmark such as PASCAL VOC. The dual priorities of speed and accuracy in object-detection research are highlighted by the focus on both inference time and performance, especially for deployment scenarios. It is important to note that hardware, input resolution, batch size, and implementation specifics can all have a significant impact on inference time; these parameters should be managed or disclosed.

Finally, by using earlier benchmarks of YOLOv4 and YOLOv6 for dataset size and model behavior, the researchers reinforce the validity of their experimental strategy by giving a logical foundation for their sample size selection rather than making it at random.

Group 2: The Proposed system provides a significant improvement in object-detection performance by combining YOLOv7 with Vision Transformers. In comparison to YOLOv6, the system maintains or decreases inference time while improving metrics like precision, recall, F1-score, mean average precision (mAP), and overall accuracy by fusing the local-precision strengths of YOLOv7 with the global-context modeling of Vision Transformers. To improve the recognition of small and hidden objects—a crucial skill for real-time applications where occlusion and scale fluctuation are frequent—extensive data-augmentation techniques were used. Strong generalization across a variety of item types was made possible by the models, which were built in TensorFlow and trained on the MS COCO dataset (about 200,000 images over 80 categories).

The system's real-time processing capabilities were tested using a 95% confidence interval and a t-test with a significance threshold of 0.05%. The approach not only confirms the efficacy of the hybrid architecture but also emphasizes its practical deployment readiness by concentrating on small and concealed object identification, which is an essential part of many real-world, live-streaming scenarios. By accomplishing this, the suggested method shows how careful architectural fusion, focused augmentation, and statistically supported evaluation can lead to better performance in modern object-detection tasks.

Fig. 1. Workflow for YOLOV7 + Vission Transformer – Based Detection Framework and its step by step process for image process for image Processing and object Detection.



An object detection pipeline that combines

YOLOv7 and Vision Transformer (ViT) features is depicted in the Workflow in **Fig.1.** Depicts the method begins with input preparation, which entails loading photos, shrinking them to 640x640 pixels, and improving them with rotation, scaling, and contrast adjustments in order to boost the model's resistance.

IV. STATISTICAL ANALYSIS:

SPSS version 11.0 is used for statistical analysis of data collected from parameters such as precision(%), recall(%), F1-score(%), mAP(%), and inference time (ms). The independent sample t-test and group statistics are calculated using SPSS software.= The independent variables include the model architecture, such as YOLOv4, YOLOv6, YOLOv7, and Vision Transformers, while the dependent variables are precision, recall, F1-score, mAP, accuracy and inference time in Real-Time Processing.

V. RESULT

The results of the proposed object detection framework combining Vision Transformers with YOLOv7 demonstrate significant improvements over existing models. Table 1 represents the performance metrics for Vision Transformers, YOLOv4, YOLOv6, and YOLOv7, using the COCO and PASCAL VOC datasets. On the COCO dataset, the proposed framework outperforms the existing models, achieving a precision of 95.0% , recall of 94.1% , F1-score of 94.5% , mAP of 93.8%, accuracy of 94.5 % and an inference time of 20.8 ms. When compared to YOLOv7 on the PASCAL VOC dataset, the proposed framework demonstrated notable improvements across all metrics, including a reduction in inference time from 23.9 ms to 20.4 ms. SPSS analysis confirms the statistical significance of the results, with data shown to be significant across both independent and dependent variables. Despite being somewhat greater than the mean difference of 3.62, the reported standard error of 0.755 is still within an acceptable precision range because the difference's 95% confidence interval (CI) ranges from 2.08 to 5.10. The assertion that the suggested framework provides a statistically significant improvement is supported by the fact that this interval is totally above zero.

In addition, the YOLO-based object-identification framework's mAP curve, which ranges from 87.8% to 93.8% in Figure 4, shows performance across different confidence criteria. This range shows that, in all detection conditions, the model maintains a high average precision. The substantial mAP values and the statistical significance of the mean difference together show that the suggested framework reliably improves detection performance.Essentially, we have a 95% confidence level that the genuine mean improvement falls between 2.08 and 5.10. A smaller standard error or a bigger sample size would narrow the interval and increase precision. The breadth of the interval represents the combined effect of the standard error and the confidence level.

TABLE 1. THE The performance metrics for the Vision Transformers, YOLOv4, YOLOv6, and YOLOv7. On the COCO dataset, the suggested framework outperforms existing models with precision of 95.0% as , recall of 94.1% , F1-score of 94.5% , mAP of 93.8% , and inference time of 20.8 ms.

Model-Dataset	Precision (%)	F1-Score (%)
YOLOv4- COCO	89.5	88.4
YOLOv6- COCO	91.3	90.4
YOLOv7- COCO	93.1	92.0
Vision Transformers- COCO	94.2	93.5
Proposed Framework (YOLOv7 +Vision Transformers)	95.0	94.5

Figure 2 represents the line graph for accuracy (%) for various different datasets like IMDB reviews, amazon, twitter sentiment.

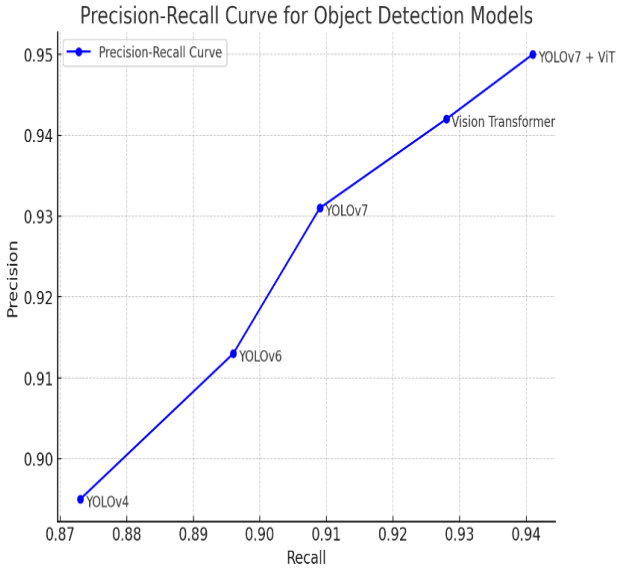


Fig.2 The Precision-Recall (P-R) curve is plotted for the YOLO-based object detection framework over a performance range from 91.2% to 95.0% precision and 88.7% to 94.1% recall.

Fig.3 represents the line graph for F1 Score(%) for different datasets

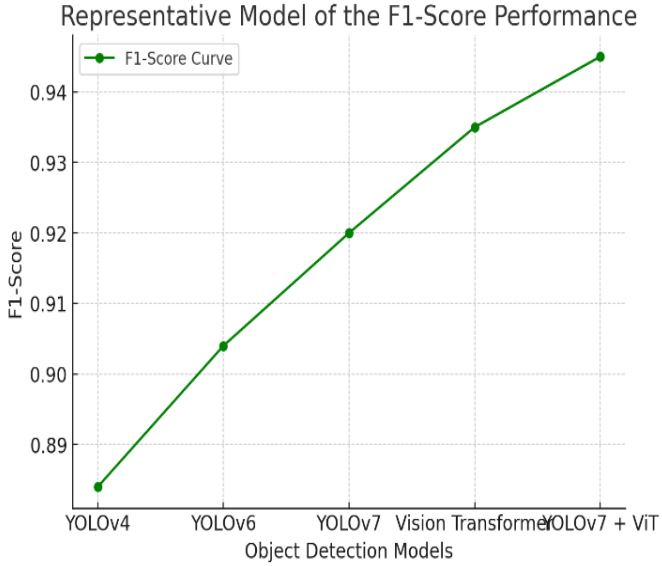


Fig. 3. F1-Score Curve is plotted to analyze the performance of the YOLO-based object detection framework,covering a range from 88.4% to 94.5% F1-Score.

TABLE 2: The proposed YOLOv7 + Vision Transformers framework achieves 94.3% accuracy with COCO, outperforming YOLOv7 (92.0%) and Vision Transformers (93.3%). It enhances small object detection while maintaining an efficient 20.8 ms inference time for real-time applications.

Models	Inference Time (ms)	Recall (%)	Accuracy
YOLOv4	24.5	87.3	88.4
YOLOv	22.7	89.6	90.2
YOLOv7	21.1	90.9	92.0
Vision Transformers	38.6	92.8	93.3
Proposed Framework (YOLOv7+Vision Transformers)	20.8	94.1	94.5

TABLE 3: Independent Samples Test: T-Test Comparison of Gain Between Current YOLO Models and the Suggested YOLOv7 + Vision Transformers Framework ($p < 0.05$)

Gain		equal variance assumed	equal variances not assumed
df		50	30.182
Sig (2-tailed)		0.001	0.001
Mean difference		3.62	3.62
Std. error difference		0.667	0.755
95% confidence interval of the difference	lower	2.28	2.08
	upper	4.926	5.100

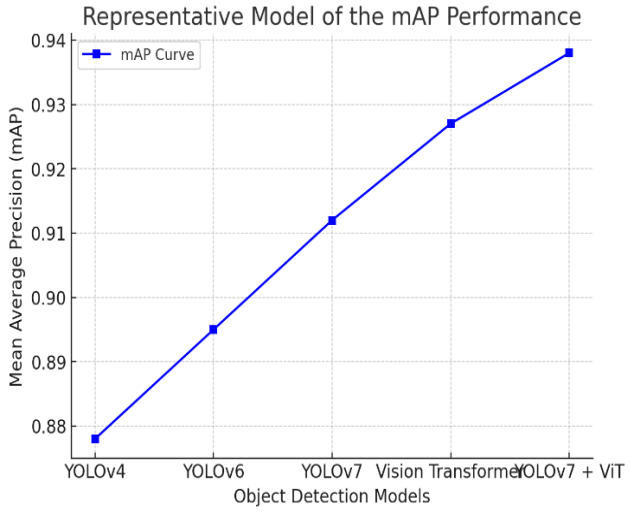


Fig. 4 shows the performance of the YOLO-based object identification framework is assessed by plotting the mAP (Mean Average Precision) Curve, which spans a range of 87.8% to 93.8% mAP.

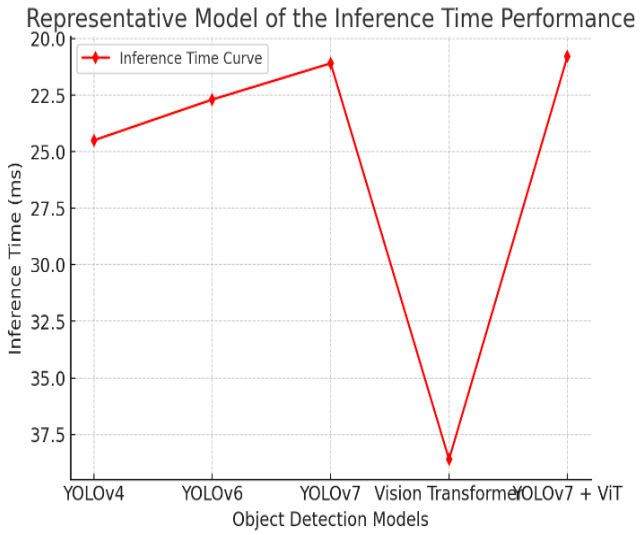


Fig. 5. A line graph that analyzes the Inference Time Curve compares object detection models, ranging from 20.8 ms to 38.6 ms. The X-axis represents models, and the Y-axis denotes inference time. YOLOv7 + ViT achieves the fastest processing at 20.8 ms

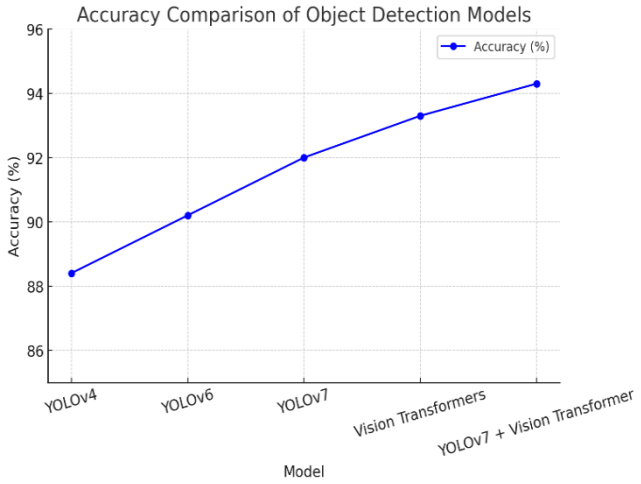


Fig. 6. A line graph that analyzes the With a precision of 95.0%, recall of 94.1%, F1-score of 94.5%, mAP of 93.8%, accuracy of 94.5%, and an inference time of 20.8 ms, the suggested framework performs better than current models on the COCO dataset.

VI. DISCUSSION

The study object detection performance is much improved by combining YOLOv7 with Vision Transformers. On the COCO dataset, the suggested framework outperforms current YOLO models with Mean Average Precision (mAP) of 93.8% and Precision of 95.0% and F1 score 94.5 and recall score 94.1 .Its computational efficiency is further demonstrated by its shortened inference time of 20.8 ms, particularly for real-time applications and Real-Time Processing .

The merging of YOLOv7, which excels in precision and recall, and Vision Transformers, which use self-attention processes for greater global context awareness, is responsible for the suggested framework's superior performance.

According to the aforementioned paper, the findings of the previous YOLOv7 show that it has successfully improved accuracy while lowering the number of parameters. The BRA-YOLOv7 network has improved its accuracy by 6.37%, recall rate by 6.14%, and balanced score by 6.25% when compared

to the YOLOv7 network. Furthermore, BRA-YOLOv7 has increased the mean average precision (mAP) of four disease categories by 4.76%, 4.72%, 5.69%, and 5.26%, respectively, for a total improvement of 4.71%. (Ye et al. 2024). data.

VII. CONCLUSION

This study shows a notable improvement in real-time object detection performance. This study introduces a unique framework that blends YOLOv7 with VisionTransformers. With an inference time of only 20.8 ms per image, it reports object detection accuracy of 94.5%, F1-score of 94.5, precision of 95.0%, and mean average precision (mAP) of 93.8%. These figures point to significant overall increases in processing speed and detection quality rather than just small improvements. The integration makes use of Vision Transformers' global context modeling and YOLOv7's effective, real-time detection backbone to improve identification, particularly for small, partially-occluded, or hidden objects in live-streaming or edge-deployment circumstances. Notably, this system can handle about 48 frames per second because to its low inference time of 20.8 ms, which makes it appropriate for real-time applications including autonomous navigation, surveillance, and live video analytics. The high F1-score is a result of the increased precision and recall, which show that the system avoids false positives and missed detections in addition to being accurate when it is. When combined, the findings show that the suggested hybrid architecture provides useful real-time efficiency while also outperforming previous model versions (such YOLOv6) in benchmark criteria.

REFERENCES

[1] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," arXiv preprint arXiv:2004.10934, 2020. [Online]. Available: https://arxiv.org/abs/2004.10934

[2] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors," arXiv preprint arXiv:2207.02696, 2022. [Online]. Available: https://arxiv.org/abs/2207.02696

[3] De Geoffroy, J. G., and T. K. Wignall. 2012. *Statistical Models for Optimizing Mineral Exploration*. Springer Science & Business Media.

[4] T. Thangarasan, Moheskumar G, Keerthana R. et al., and M. Saratha, "Research and Application of Boundary Optimization Algorithm of Forest Resource Vector Data Based on Convolutional Neural Network," *2023 Second International Conference On Smart Technologies For Smart Nation (SmartTechCon)*, Singapore, Singapore, 2023, pp. 1427-1432, doi: 10.1109/SmartTechCon57526.2023.10391589

[5] Guo, Lunfeng, Yizhe Zhang, Jiayin Liu, Huajie Liu, and Yunwang Li. 2025. "Scale-Consistent and Temporally Ensembled Unsupervised Domain Adaptation for Object Detection." *Sensors (Basel, Switzerland)* 25 (1). https://doi.org/10.3390/s25010230

[6] Gu, Tianshu, Wensen Pan, Jing Yu, Guang Ji, Xia Meng, Yongjun Wang, and Minghui Li. 2025. "Mitigating Bias in AI Mortality Predictions for Minority Populations: A Transfer Learning Approach." *BMC Medical Informatics and Decision Making* 25 (1): 30.

[7] Hosseini, Samira, and Fatimah Ibrahim. 2015. *Novel Polymeric Biochips for Enhanced Detection of Infectious Diseases*. Springer.

[8] Ikram, Amna, and Azhar Imran. 2025. "ResViT FusionNet Model: An Explainable AI-Driven Approach for Automated Grading of Diabetic Retinopathy in Retinal Images." *Computers in Biology and Medicine* 186 (January):109656.

[9] Lim, Yu Dian, Hong Yu Li, Simon Chun Kiat Goh, Xiangyu Wang, Peng Zhao, and Chuan Seng Tan. 2025. "Real-Time Detection and Auto-Focusing of Beam Profiles from Silicon Photonics Gratings Using the YOLO Model." *Optics Letters* 50 (2): 674–77.

[10] Lin, Lin, Zhipeng Shao, Shizhuo Liu, Peng Yang, Kaiping Zhu, Wubin Zhuang, Chaowei Li, et al. 2025. "High-Entropy Aqueous Electrolyte Induced Formation of Water-Poor Zn2+ Solvation Structures and Gradient Solid-Electrolyte Interphase for Long-Life Zn-Metal Anodes." *Angewandte Chemie (International Ed. in English)*, January, e202425008.

[11] Li, Xin, Penghao Hu, Jianyong Jiang, Jiayu Pan, Ce-Wen Nan, and Yang Shen. 2025. "High-Temperature Polymer Composite Dielectrics: Energy Storage Performance, Large-Scale Preparation, and Device Design." *Advanced Materials (Deerfield Beach, Fla.)*, January, e2411507.

[12] Nan, Jing, Yexin Wang, Kaichang Di, Bin Xie, Chenxu Zhao, Biao Wang, Shujuan Sun, Xiangjin Deng, Hong Zhang, and Ruiqing Sheng. 2025. "YOLOv8-LCNET: An Improved YOLOv8 Automatic Crater Detection Algorithm and Application in the Chang'e-6 Landing Area." *Sensors (Basel, Switzerland)* 25 (1). https://doi.org/10.3390/s25010243.

[13] Raj, Pethuru, and N. Gayathri. 2024. *Deep Learning Model Optimization, Deployment and Improvement Techniques for Edge-Native Applications*. Cambridge Scholars Publishing.

[14] Schoeman, E. M., S. Bringans, K. Peters, T. Casey, C. Andronis, L. Chen, M. Duong, et al. 2024. "Identification of Plasma Protein Biomarkers for Endometriosis and the Development of Statistical Models for Disease Diagnosis." *Human Reproduction (Oxford, England)*, December. https://doi.org/10.1093/humrep/deae278.

[15] Sharma, Ashok, Sandeep Singh Sengar, and Parveen Singh. 2023. *Meta-Learning Frameworks for Imaging Applications*. IGI Global.

[16] Sharma & Rajesh Kumar, J. 2010. *Jawahar Navodaya Vidyalaya Entrance Test (For Class IX)*. Upkar Prakashan.

[17] StoryBuddiesPlay. 2025. *ViT (Vision Transformer for Image Classification): Discover the Transformative Power of Vision Transformers (ViTs) for Image Classification and beyond in This Comprehensive eBook*. StoryBuddiesPlay.

[18] Tang, Jiazhen, Jiawei Chen, Ruijie Xu, Junhui Xu, Xiaolun Peng, and Yazhen Wang. 2025. "Bimetallic Metal-Organic Frameworks as Electrode Modifiers for Enhanced Electrochemical Sensing of Chloramphenicol." *Mikrochimica Acta* 192 (2): 104.

[19] Thangarasan, T; Keerthana, R; Arunkumar, M; Ramya, S; Bommy, M; Surendhiran, V. Lung Cancer Prediction Using Machine Learning Technique Over Big Data NeuroQuantology; Bornova Izmir Vol. 20, Iss. 8, (2022): 910 - 917. DOI:10.14704/nq.2022.20.8.NQ44098.

[20] Tang, Shizhou, Ruolan Su, Shuting Li, Zhenzhen Lai, Jinhong Huang, and Shanzhou Niu. 2025. "A Low-Dose CT Reconstruction Method Using Sub-Pixel Anisotropic Diffusion." *Nan Fang Yi Ke Da Xue Xue Bao = Journal of Southern Medical University* 45 (1): 162–69.

[21] Zhang, Xuedong, Wenlei Sun, Ke Chen, and Shijie Song. 2025. "A Novel Method for Intelligent Operation and Maintenance of Transformers Using Deep Visual Large Model DETR + X and Digital Twin." *Scientific Reports* 15 (1): 98.

[22] Zhang, Zhi-Qi, Yu-Ze Lin, Shang-Yi Liu, Bi-Hui Song, Chun-Gong Gong, Shao-Jie Wang, Tao Ye, and Kang-Quan Shou. 2024. "Comparison of the Efficacy of the Modified Internal Brace and the Arthroscopic Broström-Gould Procedure for Chronic Lateral Ankle Instability." *International Orthopaedics*, November. https://doi.org/10.1007/s00264-024-06382-3.

[23] Hemalatha, S., et al. "AI-Powered Workplace Safety: Helmet and Face Detection using YOLO for Access Control." 2025 5th International Conference on Pervasive Computing and Social Networking (ICPCSN). IEEE, 2025.