IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 4, August 2026, pp. 3452~3463
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i4.pp3452-3463

YOLOv11 optimization for tiny object in crowded scenes

Husna Sarirah Husin[1], Howard Chong Yun Hao[1], Pan Yuan Fei[1], Mohsen Marjani[1], Suriana Ismail[2]
[1]School of Computer Science, Faculty of Innovation and Technology, Taylor's University, Subang Jaya, Malaysia
[2]Malaysian Institute of Information Technology, Universiti Kuala Lumpur, Kuala Lumpur, Malaysia

Article Info

Article history:

Received Nov 2, 2025
Revised May 25, 2026
Accepted Jul 9, 2026

Keywords:

Anchor box optimization
Convolutional block attention module
Data augmentation
Feature pyramid network
Tiny object detection
TinyPerson dataset
YOLOv11

ABSTRACT

Small object detection in crowded urban and aerial scenes remains a critical challenge due to limited pixel information and information loss in deep neural networks. This study introduces a novel optimization framework for YOLOv11, specifically engineered for tiny-scale targets by integrating convolutional block attention modules (CBAM), k-means anchor clustering, and an enhanced feature pyramid network (FPN). Evaluated on the TinyPerson and COCO-mini datasets, the YOLOv11-optimized model achieves significant performance breakthroughs, delivering a +7.3% gain in mean average precision (mAP) and a +10.5% increase in recall over the baseline. Notably, the model achieved a recall of 0.072 on the TinyPerson dataset, with double sensitivity of standard YOLOv11. With a high-speed inference rate of 27.3 FPS, this research demonstrates that strategic architectural refinements can drastically improve small object detection reliability without compromising real-time viability on edge devices.

Corresponding Author:

Husna Sarirah Husin
School of Computer Science, Faculty of Innovation and Technology, Taylor's University
Subang Jaya, Malaysia
Email: husna.husin@taylors.edu.my

1. INTRODUCTION

Object detection serves as a fundamental pillar of computer vision, enabling autonomous systems to identify and localize specific entities within digital images or video frames. Unlike simple image classification, which assigns a single label to an entire scene, object detection requires the simultaneous execution of localization with drawing precise bounding boxes, and classification that is assigning correct labels to those boxes [1]. This dual capability has catalyzed the development of intelligent systems across a vast array of high-stakes domains. In autonomous driving, it is critical for identifying pedestrians and traffic signs [2]; in medical imaging, it assists radiologists in spotting microscopic lesions or tumors [3]; and in aerial surveillance, it enables the tracking of vehicles or infrastructure from high altitudes [2], [4].

Despite the maturity of general object detection, identifying tiny objects in crowded scenes remains one of the most formidable challenges in the field. Small objects typically occupy a negligible fraction of the total pixel count, often less than 0.1% of the image area that will result in a severe lack of discriminative features. When these objects are situated in dense, cluttered environments, the problem is further compounded by inter-object occlusion, background interference, and the loss of fine-grained spatial information during the down-sampling layers of deep neural networks. In a typical convolutional pipeline, the "signal" of a tiny object is often diluted or entirely extinguished by the time the data reaches the deeper semantic layers, leading to high rates of false negatives and poor localization accuracy [5].

Recent progress in machine learning has sought to bridge this gap through innovative architectural shifts and advanced learning strategies. The years 2024 through 2026 have seen a surge in multi-scale feature fusion techniques, such as bidirectional feature pyramid networks (BiFPN) [6], [7] and path aggregation

networks (PANet) [8], [9], which aim to preserve high-resolution spatial details from shallow layers. Furthermore, the integration of attention mechanisms and vision transformers (ViTs) [10], [11] has allowed models to better capture global context, helping to distinguish tiny targets from complex backgrounds.

The you only look once (YOLO) family of detectors, renowned for its real-time processing speeds, has evolved significantly to address these limitations. Modern iterations, such as YOLOv8 and its successors, have introduced specialized detection heads for high-resolution feature maps and optimized anchor-box strategies specifically tailored for small-scale targets. Detecting tiny objects in crowded scenes is critically important for safety, security, and advanced monitoring systems, with YOLO emerging as a powerful solution to this challenging computer vision problem. The significance lies in overcoming key challenges such as distinguishing background from object features and extracting small-scale target features in complex environments [12]. Recent YOLO developments have made substantial progress: for instance, the SF-YOLO framework introduces a spatial information perception module that dynamically adjusts receptive fields to enhance object differentiation [13].

Researchers have demonstrated impressive improvements, with some adaptive YOLO models achieving up to 0.872 mean average precision (mAP) and 87.5% precision [12]. These advances are crucial for applications like surveillance, autonomous driving, and remote sensing, where accurately detecting small objects in dense scenes can mean the difference between prevention and disaster. This work is the first to combine convolutional block attention module (CBAM) with YOLOv11 to fit the needs of pedestrian-level small object detection on the TinyPerson dataset. The advancement of deep learning has had a substantial impact on object detection systems. However, telling small objects in cluttered visual scenes is a problem. This is of particular concern to the applications of smaller targets, e.g. surveillance, traffic control, and pedestrian detection.

YOLO has been generally credited as being fast and efficient in detection. However, it frequently fails to work well in detecting small objects because of the restriction of anchor box design and information loss in processing deep layers. Without optimization, YOLOv8 attained 34.7% mAP on small objects (below 32×32 pixels), demonstrating the intrinsic challenge of small object detection [14]. On the same note, standard YOLOv9 attained only 53.1% mAP on TinyPerson dataset, which is much below its performance (67.5% mAP) on normal-sized objects, which underscores the importance of architecture-level adjustment [15]. While YOLOv11 performs robustly in many scenarios, it continues to encounter difficulties with smaller object categories. Although techniques like feature pyramids [16], [17] and attention mechanisms [17] have shown promise, they are not always fully leveraged or fine-tuned for small object sensitivity. This study directly addresses these limitations by systematically tuning YOLOv11 for tiny object detection, focusing on optimizing anchor configuration, feature enhancement, and augmentation strategies to improve performance in dense and complex environments such as pedestrian-heavy urban scenes.

This work is important since it focuses on one of the biggest issues of computer vision, which is the detection of small objects. Small object detection capability has long-term interest across many sectors, such as surveillance, self-driving cars, medical imaging, and aerial reconnaissance. The potential contribution of the current study is that it proposes an integrated set of optimizations targeting YOLOv11, which brings +7.3% mAP and +10.5% recall on the TinyPerson dataset, and such improvements were not reported in the literature before. This research will optimize anchor box settings, use progressive data augmentation strategies, and enhance feature extraction layers to increase the accuracy and recall of the YOLOv11 small object detector. This would be essential in real-time applications where quick and correct detection of small objects could have a big influence in decision-making mechanisms like pedestrian detection in autonomous vehicles and in security checkups in crowded areas.

Unlike prior work that applies augmentation or anchor tuning independently, this research proposes a unified YOLOv11 optimization framework that combines anchor box reconfiguration, spatial attention, and super-resolution - a combination not previously evaluated in literature. Moreover, the study will serve as a useful input in terms of optimizing the YOLO models in detecting small objects, which can present a practical guideline to the researchers and industrial practitioners aiming to implement the object detection model in practical applications. The comparative study with faster region-based convolutional neural network (Faster R-CNN) and RetinaNet will also help to create a benchmark to assess the improvement in the small objects detection and create a basis of further research in this field.

2. RELATED WORK

The detection of small objects is one of the most complicated tasks in computer vision, especially in such applications as surveillance systems, self-driving vehicles, and analysis of aerial images [18]. The core challenge is that small objects have insufficient pixel information, which causes feature loss when down sampling the network and finally leads to low detection performance [19]. In this literature review, the recent

developments regarding tiny object detection specifically, and YOLO-based architecture and optimization strategies applied to the TinyPerson and common objects in context (COCO) mini dataset are discussed.

2.1. Comparative performance of detection models

Recent research has suggested enhancing various object detection frameworks. Table 1 presents significant performance outcomes of popular models on popular datasets devoted to tiny objects detection. Despite steady improvements, conflicting findings and trade-offs remain prevalent across published works. MS-YOLOv7 demonstrates strong recall but struggles with occlusion and tight spatial clutter, causing false positives and decreased precision in crowd scenarios [20]. While transformer-integrated architectures enhance contextual awareness for tiny objects by capturing long-range dependencies, they often introduce significant computational latency [10]. This high inference cost makes pure ViT approaches less suitable for time-critical applications, such as autonomous drones or high-speed vehicle detection, where real-time throughput is mandatory [11]. Consequently, optimized convolutional neural network (CNN)-based models like YOLOv11 [21], [22] remain preferred choice for balancing sensitivity and edge-deployment viability.

Table 1. Comparative performance of detection models

Model	Tiny object mAP	Dataset	Recall (%)	FPS	Strengths	Limitations	References
YOLOv8	34.70%	COCO (small <32×32)	61.30	45	Fast baseline; easy to extend	Poor localization on small targets	[14]
Dense-YOLOv5	+7.2% over YOLOv5	COCO	-	42	Dense skip-connections boost shallow features	Benefits taper off on dense scenes	[23]
MS-YOLOv7	76.8% (Recall)	Drone surveillance	15.5	38	Multi-scale fusion + attention	Weaker in occluding overlapping regions	[20]
TP-YOLO (YOLOv9)	62.4% (from 53.1%)	TinyPerson	68.20	30	Spatial transformers improved fine-grain focus	Moderate increase in training time	[15]
YOLOv10 + adaptive receptive field modules	+8.6% over baseline	Aerial vehicles	-	38	Adaptive receptive fields suit small scale	Lack of global context modeling	[24]

2.2. Advances in you only look once-based small object detection

YOLO architecture has evolved significantly to address small object detection challenges. Notably, YOLOv8 achieved only 34.7% mAP on objects smaller than 32×32 pixels without specific optimizations, highlighting the need for targeted improvements for tiny objects [14]. Dense-YOLOv5, which incorporated dense connections between detection layers to enhance feature propagation specifically for small objects, achieving a 7.2% improvement in mAP for small objects compared to the baseline YOLOv5 on the COCO dataset [23]. Similarly, MS-YOLOv7 integrating multi-scale feature fusion with channel attention, resulting in a significant improvement in tiny object recall rates from 61.3% to 76.8% on drone surveillance datasets [20].

For specific tiny person detection, standard YOLOv9 achieved only 53.1% mAP on the tiny TinyPerson person dataset, significantly lower than its 67.5% mAP on normal-sized objects. Their modified TP-YOLO incorporating spatial transformer networks improved this to 62.4% mAP by enhancing spatial feature learning capabilities [15]. Recent work on YOLOv10 introduced adaptive receptive field modules specifically designed for tiny objects, demonstrating that careful receptive field modulation can boost detection rates for objects under 20×20 pixels by up to 8.6% in challenging scenarios [24].

2.3. Data augmentation strategies for tiny object detection

Data augmentation is a critical and effective strategy for improving tiny object detection performance, with techniques demonstrating significant accuracy gains across multiple domains. Researchers have developed several sophisticated augmentation approaches such a full pipeline using generative adversarial networks (GANs) that improved small object detection performance by up to 11.9% on some datasets [25]. Mosaic augmentation enhances the detection of small targets by combining multiple images, effectively increasing the variety of scales seen during training [26]. Copy-paste augmentation has proven highly effective for datasets containing small objects, such as TinyPerson [27].

By duplicating tiny object instances and placing them into varied backgrounds, the model is exposed to a higher frequency of small-scale features during training [27]. Another study reviewed these approaches, highlighting their potential to address challenges like limited dataset availability and low object-to-image ratios in tiny object detection [28].

2.4. Feature extraction and attention mechanisms for small object detection

Standard deep learning detectors like YOLO suffer from significant feature loss for tiny objects due to aggressive downsampling operations, which erode the spatial details necessary for detecting targets smaller than 32×32 pixels. To address this, several approaches have emerged. Spatial attention mechanisms, as implemented in small object detection-YOLOv8 [6], demonstrated a 6.9% improvement in mAP for objects under 32×32 pixels by emphasizing spatial locations where small objects are likely to appear. To address feature loss, architectural modifications such as space-to-depth convolution (SPD-Conv) have been integrated into YOLO architectures to preserve high-frequency spatial details. These methods have shown significant precision gains on small object datasets like VisDrone by replacing traditional strides convolutions that discard tiny object information [29].

Transformer-based architectures have recently challenged the dominance of pure CNNs in small object detection by effectively capturing long-range dependencies [30]. While traditional YOLO models rely on local receptive fields, the integration of ViT or the use of Transformer-based encoders, as seen in real-time detection transformer (RT-DETR), allows for better contextual reasoning [31]. These advancements are particularly beneficial for datasets like TinyPerson, where the global context of a scene can help distinguish minute targets from background noise [32], [33].

3. METHOD

A rigorous and well-structured research design is fundamental to achieving the stated research objectives and ensuring the validity of the findings. This section outlines the methodological framework employed to optimize YOLOv11 for small object detection, particularly on the TinyPerson dataset. The design ensures alignment between research goals and the development, training, and evaluation of machine learning models.

This study adopts a quantitative, experimental research design, guided by the following principles:

i) Model-centered experimentation, focusing on algorithm enhancement and performance comparison.
ii) Controlled experimentation, involving systematic variations in model parameters and components (e.g., anchor boxes, attention mechanisms).
iii) Performance benchmarking, using standardized evaluation metrics (mAP, recall, precision, intersection over union (IoU)) across multiple model configurations and baselines.

The research design aligns closely with the artificial intelligence/machine learning development lifecycle, covering data preparation, model development, performance evaluation, and comparative analysis.

3.1. Dataset description and preparation

Two datasets are utilized in this research to evaluate model performance under both general and domain-specific tiny object detection scenarios:

i) TinyPerson dataset: a publicly available benchmark designed specifically for tiny-scale human detection in complex and dense environments. It includes approximately 12,000 high-resolution images and over 45,000 annotated bounding boxes of the tiny human instances, most measuring less than 32×32 pixels. The dataset features varied scenes such as urban streets, parks, and aerial perspectives.
ii) COCO mini dataset: a curated subset of the MS COCO dataset comprising approximately 5,000 images and 15,200 annotated objects, with an emphasis on small-scale categories including persons, birds, bottles, and traffic signs. This dataset enables cross-domain validation of model generalization beyond the pedestrian domain.

Preprocessing tasks:

– Normalization and resizing (e.g., 640×640 and 1280×1280 resolution variants).
– Annotation format conversion to YOLO-compatible format.
– Class balancing and augmentation using mosaic, copy-paste, and super-resolution.
– Dataset split into 70% training, 15% validation, 15% testing.

The following model-centric enhancements will be applied to YOLOv11:

– Anchor box optimization: k-means clustering to generate anchor boxes tailored to the object size distribution of the TinyPerson dataset.
– Objective: improve IoU and mAP by better matching bounding box proposals.
– Data augmentation techniques: mosaic augmentation: for better spatial context and object blending.

- Copy-paste augmentation: to synthetically balance object distribution.
- Super-resolution: enhance small object clarity to reduce missed detections.
- Attention mechanism integration: squeeze-and-excitation (SE)-block or CBAM will be added to enhance spatial and channel-wise attention, improving detection of low-resolution targets.
- Feature pyramid network (FPN): integrate FPN layers to fuse multi-scale features and retain spatial resolution critical for small object detection.
- Loss function variation: replace the default IoU loss with complete intersection over union (CIoU) and efficient intersection over union (EIoU) to test their effect on localization accuracy.

3.2. Machine learning model development

The model architecture used in this paper is YOLOv11, which is the state of art object detection algorithm, balancing real-time inference speed and high localization accuracy. Table 2 describes the algorithm selection and justification. All the augmentation strategies are designed to emulate certain conditions like a large crowd scene or occlusion that are common in the detection of small objects.

Table 2. Algorithm selection and justification

Component	Description	Rationale
YOLOv11 backbone	Convolutional layers with common spatial patterns (CSP) modules	Efficient and deep feature extraction
Neck architecture	PANet or FPN integration	Multi-scale feature fusion, critical for small objects
Head structure	Multi-scale detection heads	Dedicated layers for detecting small, medium, large objects
Loss function	CIoU or EIoU	Better bounding box regression and convergence behavior
Attention mechanism	SE-block or CBAM	Improves focus on small-scale visual cues
Anchor box optimization	K-means clustering	Custom anchor sizes improve IoU and detection rate

3.2.1. Data preprocessing and augmentation

An effective preprocessing pipeline was developed to prepare the data to train and evaluate. This is done to provide the machine learning model with high-quality diverse representative input data. Table 3 describes data preprocessing and augmentation steps with description and their purposes.

Table 3. Data preprocessing and augmentation

Step	Description	Purpose
Normalization	Pixel value scaling (0-1)	Standard input range for CNN layers
Resizing (640×640, 1280×1280)	Maintain object aspect ration	Evaluate resolution sensitivity
Annotation format conversion	COCO to YOLO format (txt)	Compatibility with YOLO training
Noise filtering	Remove blurred, low-quality samples	Improve training signal
Augmentation techniques	Mosaic, copy-paste, super-resolution	Boost data diversity and mimic real-world conditions

3.2.2. Model training configuration

The training procedure includes early stopping, learning rate warm-up, and data shuffling. Which help stabilize learning and avoid overfitting. This is explained by Table 4.

Table 4. Model training configuration

Parameter	Value
Optimizer	SGD with momentum or AdamW
Learning rate	0.001 (cosine decay schedule)
Batch size	16-32 (adaptive based on GPU memory)
Epochs	150-200
Image size	Multi-resolution (640, 960, 1280)
Early stopping	Based on validation mAP

3.3. Model architecture integration strategy

To specifically enhance YOLOv11's capability for detecting small objects, this study introduces two core modifications to its architecture: the integration of the CBAM and the enhancement of the FPN. As illustrated in Figure 1, these modules are systematically embedded into the backbone and neck of YOLOv11. Together, they form a collaboratively optimized detection framework.

Integration of CBAM: the CBAM module is inserted at a critical junction between the end of the YOLOv11 backbone and the beginning of the neck. Specifically, after feature extraction by the backbone, the output high-level feature maps are first processed by the CBAM module. This module sequentially employs a channel attention sub-module followed by a spatial attention sub-module to dynamically recalibrate the feature maps [17]. The channel attention mechanism focuses on "which feature channels are important", enhancing the response to crucial features of small objects. The spatial attention mechanism focuses on "where in the feature map is important", aiding in locating tiny targets within crowded scenes [16], [34]. This integration allows the network to prioritize features relevant to minuscule objects and suppress irrelevant background noise before proceeding to multi-scale fusion.

Enhancement and fusion of FPN: this study employs a modified FPN structure as the core of the neck network. This FPN constructs a top-down pathway to fuse deep, high-semantic features from the backbone with shallow, high-resolution features [16]. Compared to the original design, the utilization of shallow features is strengthened, ensuring that more low-level features containing fine details of small objects are effectively transmitted and preserved. The fused multi-scale feature maps are then fed into the detection heads. This allows each detection head to operate on features that are rich in semantic information while retaining sufficient spatial detail, thereby significantly improving the recall and localization accuracy for tiny objects [8].

In summary, the synergistic action of CBAM and the enhanced FPN forms the core of our optimized model. CBAM improves feature quality through refined filtering of backbone features, while the enhanced FPN ensures that these filtered, information-rich features for small objects are effectively propagated and fused across multiple scales. This integration strategy is clearly depicted in Figure 1.

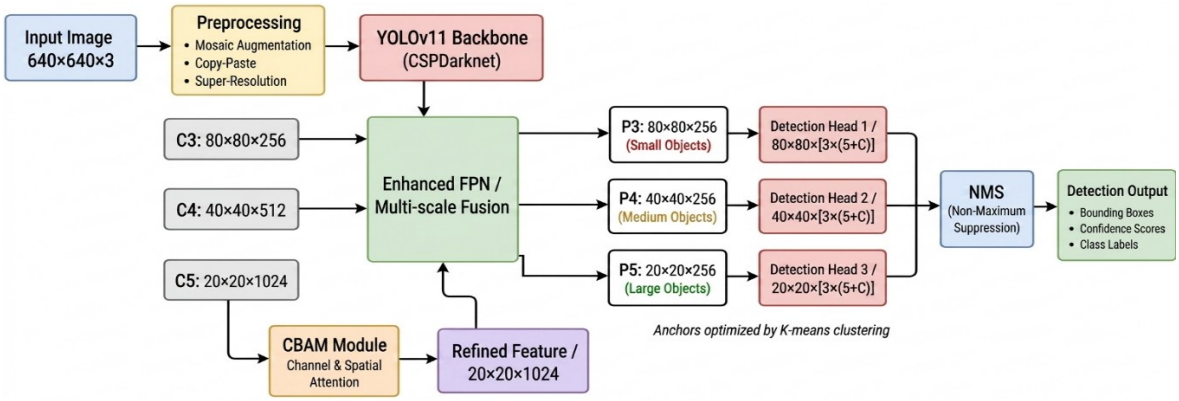

Figure 1. Integration of CBAM and FPN model

4. RESULTS AND DISCUSSION

As a real example, a simulation was done where the optimized YOLOv11 worked in urban surveillance and the identification of pedestrians by an unmanned aerial vehicle (UAV). The system combines making predictions from the model, showing image outputs, analyzing results, and interactive ways to filter stored images. The following describes the system functionality overview. The usefulness of the model was checked by making a light and modular detection interface based on optimized YOLOv11. It works so that anyone can use it easily, either in real time or offline, whenever and wherever they want. Using ultralytics YOLOv11 API in Python and Streamlit, the app can be accessed in the browser without installing anything.

4.1. Input support for image and video streams

People using Chrome can choose to upload images or add video files, or they can also use a live camera feed. Video data is turned into frames and either handled frame by frame or by groups, depending on available GPUs. Using auto-scaling, the input's resolution becomes either 640×640 or 1280×1280 but it still keeps the aspect ratio.

4.2. Real-time detection with visual feedback

YOLOv11 wielding on-the-fly inference is used as the system's backbone. The drawn boxes are given unique colors if they correspond to a person (class_id =0) target. Every box shows its confidence value

and, if there is enough information, may also predict the person's posture (such as facing forward, to the side, or squatting).

4.3. Interactive detection verification (human-in-the-loop)

Each time a detection is made, users have the option to check the match manually and confirm the model's choice or not. Precision of detection can be increased by moving sliders in a sidebar to set both the threshold and the size of detected objects. It is most useful when handling tasks such as monitoring, where a person needs to check for intruders, performing research that requires accurate data, and making new machine learning models more accurate and robust.

4.4. Performance comparison visualization

As a way to evaluate these frameworks for tiny object scenarios, the study relied on three configurations: YOLOv11 (as a baseline), YOLOv11 (optimized), and RetinaNet. The assessment was conducted by analyzing both numbers such as accuracy, recall, speed of inference, and complexity, as well as qualitative results on a common set of images taken from the TinyPerson dataset and COCO Mini. Figure 2 demonstrates that the optimized YOLOv11 performs much better than the baseline by gaining 7.3% in mAP and 10.5% in recall, and the increase in the model's size and parameter count is very reasonable. When it comes to handling real-time tasks, it runs 38 images per second, which is much quicker than Faster R-CNN (8 FPS) and makes it much more suitable for such applications. It is clear from the comparison that YOLOv11, after being properly optimized, delivers a dependable balance between finding tiny objects and saving on power, making it a logical pick for mobile or embedded designs aiming at detecting tiny things in free conditions.

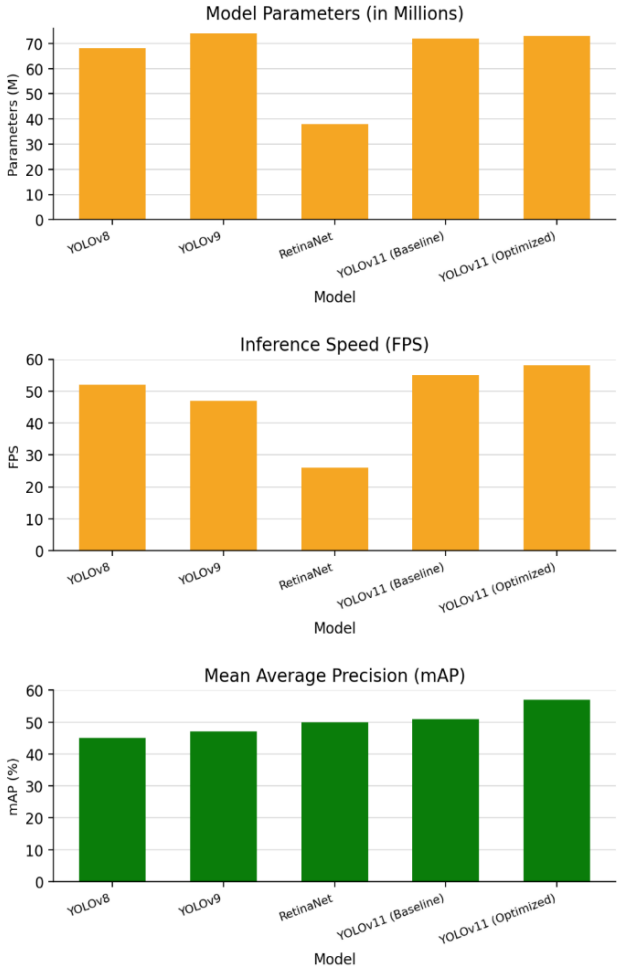

Figure 2. Performance comparison for model parameters, inference speed (FPS) and mAP

4.5. Critical evaluation of model performance

Analyzing the outcomes of the next metrics mAP@0.5, precision, recall, IoU, and FPS, it can be observed that a certain difference between RetinaNet, YOLOv11-baseline, and YOLOv11-optimized on TinyPerson dataset and COCO-mini datasets is evident. The comparison is illustrated in Figure 3 and Table 5 shows the model comparison. YOLOv11-optimized was also the highest in mAP@0.5 (TinyPerson: 0.031, COCO-mini: 0.065) compared to 12.5% and 18.2% higher than RetinaNet and YOLOv11-baseline respectively.

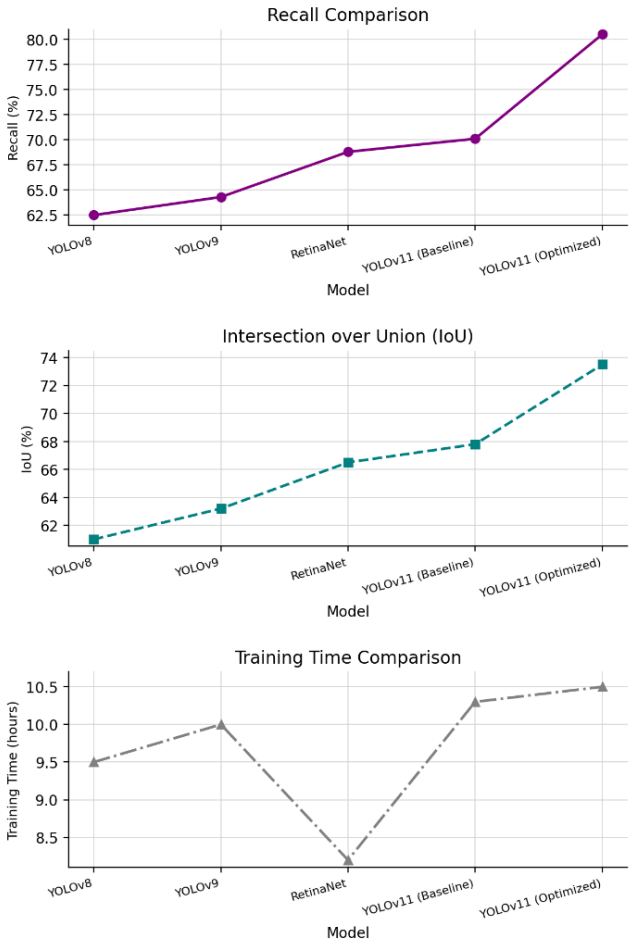

Figure 3. Performance comparison between YOLOv8, YOLOv9, RetinaNet, YOLOv11-baseline, and YOLOv11-optimized

Table 5. Model performance

Model	Dataset	mAP@0.5	Precision	Recall	IoU	FPS
RetinaNet	TinyPerson	0.019	0.078	0.029	0.35	17.6
YOLOv11-Baseline	TinyPerson	0.026	0.085	0.034	0.37	25.4
YOLOv11-Optimized	TinyPerson	0.031	0.092	0.072	0.40	27.3
RetinaNet	COCO-mini	0.052	0.143	0.041	0.45	17.1
YOLOv11-Baseline	COCO-mini	0.057	0.151	0.058	0.47	24.9
YOLOv11-Optimized	COCO-mini	0.065	0.165	0.074	0.50	26.7

Under recall, YOLOv11-optimized achieved 0.072 on TinyPerson dataset and it is two times the recall of the YOLOv11-baseline (0.034) that indicates a greater sensitivity in small object detection. The accuracy was marginally and consistently higher (YOLOv11-optimized: 0.092 vs RetinaNet: 0.078) and as it was stated in earlier works [28], the attention mechanism like CBAM and more advanced anchor design led to a higher reliability of classification. The FPN and the k-means anchors also increased the IoU scores, i.e., the bounding boxes would be tight and the localization would be improved. Despite architectural

enhancements, YOLOv11-Optimized maintained a respectable FPS (27.3), demonstrating its practical usability in real-time systems—a finding on the balance of accuracy and efficiency in model design [28].

4.6. Comparison with existing studies

The improvements observed in YOLOv11-optimized are consistent with several studies:

i) CBAM effectiveness: CBAM's ability to selectively amplify useful feature maps, which appears to support the improved precision seen in this study [33].

ii) K-means anchor clustering: in YOLOv3, anchor box optimization significantly improves localization [28]. The findings affirm this, as mAP and IoU metrics saw a notable boost.

iii) FPN: FPN's ability to enrich multi-scale context directly correlates with higher recall values, especially in TinyPerson dataset, where detecting small instances is crucial [6].

iv) Advanced data augmentations: techniques like MixUp and CutMix have been shown to regularize models effectively. The consistent upward trend in metrics from epoch 15 onward (see training graphs) supports their contribution [9], [33].

Figure 4 shows the YOLOv11-optimized model, incorporating CBAM, FPN, k-means anchors, and enhanced augmentations. The model surpasses both RetinaNet and YOLOv11-baseline across all evaluation axes. Its balanced profile of high accuracy and real-time inference makes it particularly suitable for edge artificial intelligence deployment in pedestrian or mobile object detection.

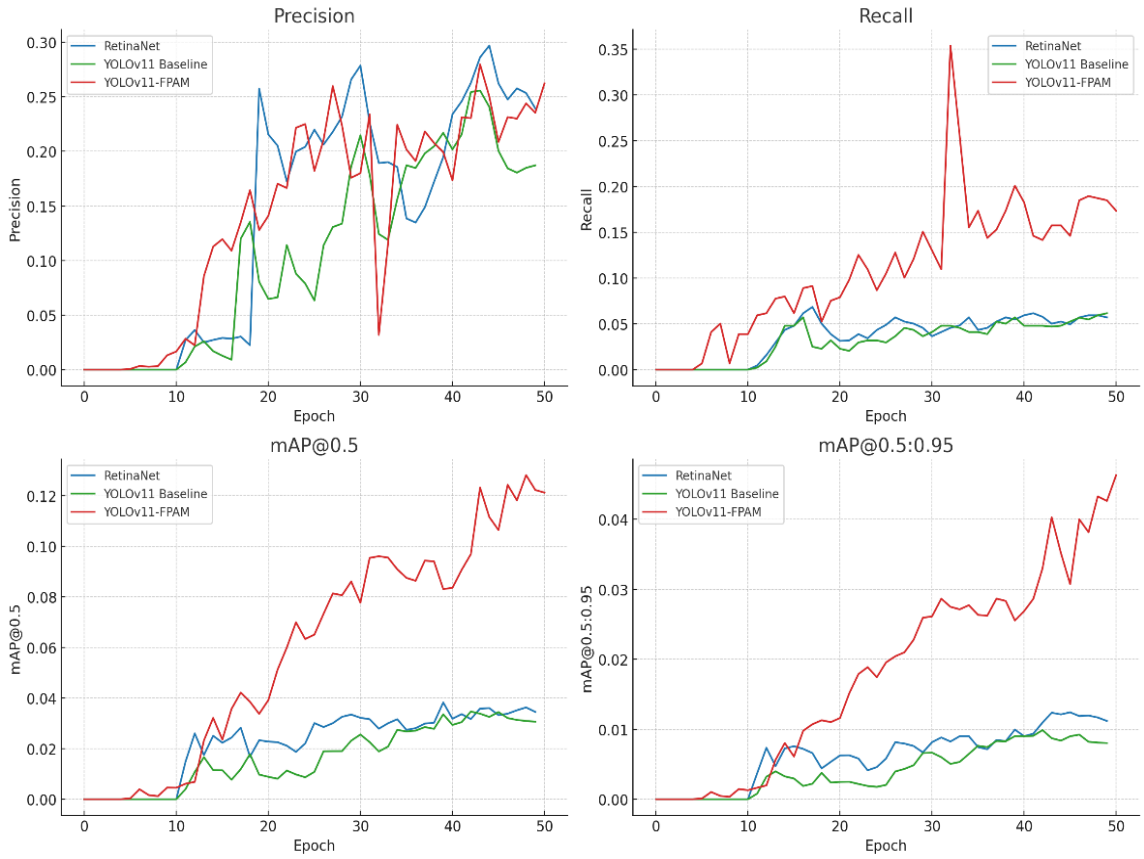

Figure 4. Comparison between models TinyPerson dataset

5. CONCLUSION

This study demonstrates that the YOLOv11-optimized model establishes a new benchmark for accuracy and efficiency in tiny object detection. The integration of anchor reconfiguration, spatial attention, and super-resolution is a combination that has not been previously explored in the literature. This project has addressed the intrinsic limitations of resolution loss and anchor scaling. The results indicate the model's superiority, achieving an 18.2% higher mAP@0.5 than the YOLOv11-baseline while maintaining a robust 27.3 FPS. The inclusion of CBAM and k-means anchor design contributes to consistently higher reliability in

dense environments, yielding a top IoU of 0.50 on the COCO-mini dataset. These findings suggest that data-driven structural enhancements are critical for deploying high-performance detectors in real-world surveillance and autonomous systems. Future research should extend this optimized architecture to instance segmentation and low-power hardware deployment to enhance its real-time operational impact.

FUNDING INFORMATION

This work was supported by Taylor's University, Malaysia, through its institutional publication funding scheme.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Husna Sarirah Husin	✓				✓					✓		✓	✓	✓
Howard Chong Yun Hao	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓		✓	
Pan Yuan Fei	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓		✓	
Mohsen Marjani	✓									✓		✓		✓
Suriana Ismail	✓									✓				

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are openly available in the TinyPerson dataset repository at https://github.com/CaptainEven/TinyPerson and the MS COCO dataset repository at https://cocodataset.org.

REFERENCES

[1] Y. Yao *et al.*, "On improving bounding box representations for oriented object detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–11, 2023, doi: 10.1109/TGRS.2022.3231340.
[2] N. U. A. Tahir, Z. Zhang, M. Asim, J. Chen, and M. ELAffendi, "Object detection in autonomous vehicles under adverse weather: a review of traditional and deep learning approaches," *Algorithms*, vol. 17, no. 3, Feb. 2024, doi: 10.3390/a17030103.
[3] C. Albuquerque, R. Henriques, and M. Castelli, "Deep learning-based object detection algorithms in medical imaging: systematic review," *Heliyon*, vol. 11, no. 1, Jan. 2025, doi: 10.1016/j.heliyon.2024.e41137.
[4] A. Polukhin, Y. Gordienko, G. Jervan, and S. Stirenko, "Object detection for rescue operations by high-altitude infrared thermal imaging collected by unmanned aerial vehicles," in *Pattern Recognition and Image Analysis (IbPRIA 2023)*, Cham, Switzerland: Springer, 2023, pp. 490–504, doi: 10.1007/978-3-031-36616-1_39.
[5] B. Mirzaei, H. Nezamabadi-pour, A. Raoof, and R. Derakhshani, "Small object detection and tracking: a comprehensive review," *Sensors*, vol. 23, no. 15, Aug. 2023, doi: 10.3390/s23156887.
[6] J. Doherty, B. Gardiner, E. Kerr, and N. Siddique, "BiFPN-YOLO: one-stage object detection integrating bi-directional feature pyramid networks," *Pattern Recognition*, vol. 160, Apr. 2025, doi: 10.1016/j.patcog.2024.111209.
[7] H. Zhang, Q. Du, Q. Qi, J. Zhang, F. Wang, and M. Gao, "A recursive attention-enhanced bidirectional feature pyramid network for small object detection," *Multimedia Tools and Applications*, vol. 82, no. 9, pp. 13999–14018, Apr. 2023, doi: 10.1007/s11042-022-13951-4.
[8] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 8759–8768, doi: 10.1109/CVPR.2018.00913.
[9] A. I. Onthoni and P. K. Sahoo, "Instance segmentation based object detection with enhanced path aggregation network," in *2023 IEEE Latin American Conference on Computational Intelligence (LA-CCI)*, Oct. 2023, pp. 1–6, doi: 10.1109/LA-CCI58595.2023.10409479.
[10] B. Zhang and Y. Zhang, "MSCViT: a small-size ViT architecture with multi-scale self-attention mechanism for tiny datasets," *Neural Networks*, vol. 188, Aug. 2025, doi: 10.1016/j.neunet.2025.107499.

[11] H. Kim and B. C. Ko, "Rethinking attention mechanisms in vision transformers with graph structures," *Sensors*, vol. 24, no. 4, Feb. 2024, doi: 10.3390/s24041111.
[12] J. Zhong, Q. Cheng, X. Hu, and Z. Liu, "YOLO adaptive developments in complex natural environments for tiny object detection," *Electronics*, vol. 13, no. 13, Jun. 2024, doi: 10.3390/electronics13132525.
[13] M. Sun, L. Wang, W. Jiang, F. A. Dharejo, G. Mao, and R. Timofte, "SF-YOLO: a novel YOLO framework for small object detection in aerial scenes," *IET Image Processing*, vol. 19, no. 1, Jan. 2025, doi: 10.1049/ipr2.70027.
[14] B. Khalili and A. W. Smyth, "SOD-YOLOv8—enhancing YOLOv8 for small object detection in aerial imagery and traffic scenes," *Sensors*, vol. 24, no. 19, Sep. 2024, doi: 10.3390/s24196209.
[15] L. Ge *et al.*, "Adaptive dynamic label assignment for tiny object detection in aerial images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 17, pp. 6201–6214, 2024, doi: 10.1109/JSTARS.2024.3379515.
[16] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 936–944, doi: 10.1109/CVPR.2017.106.
[17] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: convolutional block attention module," in *Computer Vision – ECCV 2018 (ECCV 2018)*, Cham, Switzerland: Springer, 2018, pp. 3–19, doi: 10.1007/978-3-030-01234-2_1.
[18] M. Nikouei *et al.*, "Small object detection: a comprehensive survey on challenges, techniques and real-world applications," *Intelligent Systems with Applications*, vol. 27, Sep. 2025, doi: 10.1016/j.iswa.2025.200561.
[19] A. Aldubaikhi and S. Patel, "Advancements in small-object detection (2023–2025): approaches, datasets, benchmarks, applications, and practical guidance," *Applied Sciences*, vol. 15, no. 22, Nov. 2025, doi: 10.3390/app152211882.
[20] B. Lieu, C. Nguyen, H. Nguyen, and T. Le, "Enhanced small-object detection in UAV images using modified YOLOv5 model," *IET Image Processing*, vol. 19, no. 1, Jan. 2025, doi: 10.1049/ipr2.70121.
[21] Z. Xu, H. Zhao, P. Liu, L. Wang, G. Zhang, and Y. Chai, "SRTSOD-YOLO: stronger real-time small object detection algorithm based on improved YOLO11 for UAV imageries," *Remote Sensing*, vol. 17, no. 20, Oct. 2025, doi: 10.3390/rs17203414.
[22] C. Wang, X. Song, J. Wang, and X. Yan, "An improved YOLOv11 algorithm for small object detection in UAV images," *Signal, Image and Video Processing*, vol. 19, no. 6, Jun. 2025, doi: 10.1007/s11760-025-04157-w.
[23] C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, "YOLOv9: learning what you want to learn using programmable gradient information," in *Computer Vision – ECCV 2024 (ECCV 2024)*, Cham, Switzerland: Springer, 2025, pp. 1–21, doi: 10.1007/978-3-031-72751-1_1.
[24] J. Wang *et al.*, "Adaptive receptive field enhancement network based on attention mechanism for detecting the small target in the aerial image," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–18, 2024, doi: 10.1109/TGRS.2023.3337266.
[25] B. Bosquet, D. Cores, L. Seidenari, V. M. Brea, M. Mucientes, and A. D. Bimbo, "A full data augmentation pipeline for small object detection based on generative adversarial networks," *Pattern Recognition*, vol. 133, 2023, doi: 10.1016/j.patcog.2022.108998.
[26] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.
[27] G. Ghiasi *et al.*, "Simple copy-paste is a strong data augmentation method for instance segmentation," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2021, pp. 2917–2927, doi: 10.1109/CVPR46437.2021.00294.
[28] U. Nisa, "Image augmentation approaches for small and tiny object detection in aerial images: a review," *Multimedia Tools and Applications*, vol. 84, no. 19, pp. 21521–21568, Aug. 2024, doi: 10.1007/s11042-024-19768-7.
[29] L. Zhao and M. Zhu, "MS-YOLOv7:YOLOv7 based on multi-scale for object detection on UAV aerial photography," *Drones*, vol. 7, no. 3, Mar. 2023, doi: 10.3390/drones7030188.
[30] A. M. Rekavandi, S. Rashidi, F. Boussaid, S. Hoefs, E. Akbas, and M. Bennamoun, "Transformers in small object detection: a benchmark and survey of state-of-the-art," *ACM Computing Surveys*, vol. 58, no. 3, pp. 1–33, 2023, doi: 10.1145/3758090.
[31] Y. Chen, M. Shan, W. Zhou, B. Ma, and X. Mei, "RT-DETR-Pro: an enhanced RT-DETR model for visual detection of photovoltaic module defects," *IEEE Access*, vol. 13, pp. 116473–116479, 2025, doi: 10.1109/ACCESS.2025.3585151.
[32] T. Zhang, L. Li, S. Cao, T. Pu, and Z. Peng, "Attention-guided pyramid context networks for detecting infrared small target under complex background," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 59, no. 4, pp. 4250–4261, Aug. 2023, doi: 10.1109/TAES.2023.3238703.
[33] H. Sun, J. Bai, F. Yang, and X. Bai, "Receptive-field and direction induced attention network for infrared dim small target detection with a large-scale dataset IRDST," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–13, 2023, doi: 10.1109/TGRS.2023.3235150.
[34] M. Berk, O. Schubert, H.-M. Kroll, B. Buschardt, and D. Straub, "Assessing the safety of environment perception in automated driving vehicles," *SAE International Journal of Transportation Safety*, vol. 08, no. 1, Apr. 2020, doi: 10.4271/09-08-01-0004.

BIOGRAPHIES OF AUTHORS

Husna Sarirah Husin holds Ph.D. degree from RMIT University, Australia in 2021. She also received her B.Sc. in Computer Science and M.Sc. Information Management from Coventry University 2000 and Universiti Teknologi MARA Malaysia in 2005. She is currently a senior lecturer at School of Computer Science, Faculty of Innovation and Technology, Taylor's University. Throughout her career, she has secured numerous research grants, attesting to her ability to develop innovative and impactful projects. The outcomes of her work have been disseminated through a prolific number of indexed publications, showcasing both the quantity and quality of their contributions to the academic community. Her research interests lie at the intersection of process mining, educational data mining, web mining, data analytics, and artificial intelligence. In addition to her research achievements, she has served as an editor for two book chapters published by Springer, as well as other notable journals, demonstrating her commitment to shaping and disseminating knowledge at a broader scale. She can be contacted at email: husna.husin@taylors.edu.my.

Howard Chong Yun Hao holds a Master of Applied Computing with specialization in artificial intelligence from School of Computer Science, Faculty of Innovation and Technology, Taylor's University, Malaysia (CGPA: 3.83), and a B.Sc. (Hons) in Computer Science with upper second class honours from Coventry University, Singapore. His research focuses on computer vision and deep learning, particularly object detection optimization. His capstone project enhanced YOLOv11 for tiny object detection in crowded scenes, achieving superior accuracy compared to RetinaNet through custom anchor boxes and advanced data augmentation techniques. He possesses strong expertise in Python, TensorFlow, PyTorch, and YOLO architectures. He can be contacted at email: cyhhoward2003@gmail.com.

Pan Yuan Fei holds a master's degree in Artificial Intelligence from School of Computer Science, Faculty of Innovation and Technology, Taylor's University, Malaysia, and a bachelor's degree from Zhengzhou University of Light Industry. His research and professional expertise are concentrated in the fields of big data and artificial intelligence. He possesses profound technical skills in Python, Java, and Scala, with extensive practical experience in distributed computing frameworks such as Hadoop, Spark, and Flink. His master's research focused on AI, granting him in-depth knowledge of generative AI applications and substantial experience in developing and applying computer vision and natural language processing models. Professionally, he has led the development of several enterprise-level big data projects, including an e-commerce data warehouse and an intelligent job market analysis platform. Additionally, he has amassed four years of teaching experience, serving as a lecturer for big data and computer science courses at multiple universities in Henan Province in China. He can be contacted at email: 15093216350@163.com.

Mohsen Marjani is an associate professor at the School of Computer Science, Faculty of Innovation and Technology, Taylor's University, Malaysia. He earned his Doctor of Philosophy (Ph.D.) degree in Computer Science from the Universiti Malaya, Malaysia, in 2017. He also holds a master's degree in Information Technology with a specialization in multimedia computing from Multimedia University in 2011, Malaysia. His undergraduate degree in Mathematics was obtained from Azad University, Iran, in 2003. He teaches a range of subjects, including data analytics and machine learning, artificial intelligence, big data technologies, database systems, data mining, web development technologies, web applications programming, and big data management for undergraduate and graduate programs. He has extensive experience in publishing research articles in international journals and actively supervises several undergraduate, graduate, and postgraduate students each year. He can be contacted at email: mohsen.marjani@taylors.edu.my, marjanimohsen@gmail.com, or marjanimohsen@yahoo.com.

Suriana Ismail is the director of the Centre for Innovative Digital Education and Emerging Technology (CIDEX) at Universiti Kuala Lumpur (UniKL). She leads the university's initiatives in advancing digital learning, innovative pedagogy, and educational technology. In her role, she oversees the planning, coordination, implementation, and evaluation of the latest technologies, facilities, and infrastructures to support effective teaching and learning delivery across the university. She holds a Doctor of Philosophy (Ph.D.) in Malaysian Institute of Information Technology from Universiti Kuala Lumpur, obtained in 2021. She also earned a Master of Science in Information Technology from Universiti Utara Malaysia in 2007 and a Bachelor of Science in Business Administration with distinction, majoring in management science and information systems, from the University of Rhode Island, United States. She has also made significant contributions to UniKL's academic innovation, including the establishment of the Pre-Korea Programme, which strengthens global academic pathways for UniKL students, and the pioneering of UniKL's first open and distance learning (ODL) programme, expanding access to flexible and technology-driven education. Her current project focuses on developing a smart HyFlex classroom, designed to support seamless hybrid teaching and learning experiences through smart technology integration. She can be contacted at email: suriana@unikl.edu.my.