An Efficient Face Detection and Gender Classification Approach Integrating the Speed of YOLOv9 with the Accuracy of ResNet50

Aseil Nadhim Kadhim
Faculty of Artificial Intelligence, Universiti Teknologi Malaysia, Kuala Lumpur, Malaysia
n-20@graduate.utm.my (corresponding author)

Syahid Anuar
Faculty of Artificial Intelligence, Universiti Teknologi Malaysia, Kuala Lumpur, Malaysia
syahid.anuar@utm.my

Saiful Adli Bin Ismail
Faculty of Artificial Intelligence, Universiti Teknologi Malaysia, Kuala Lumpur, Malaysia
saifuladli@utm.my

Received: 29 April 2025 | Revised: 17 June 2025, 28 June 2025, and 6 July 2025 | Accepted: 8 July 2025

 DOI: https://doi.org/10.48084/etasr.11830

ABSTRACT

Artificial Intelligence (AI), particularly deep learning models, plays a pivotal role in tasks such as face detection and gender classification. This study aims to enhance the performance of computer vision systems by developing a hybrid model that integrates YOLOv9 with a modified version of ResNet50, addressing the common trade-off between accuracy and inference speed found in traditional approaches. To achieve this, a custom dataset was collected from real-world conditions within a university campus, testing the performance of multiple You Only Look Once (YOLO) models and Convolutional Neural Network (CNN) architectures. Experimental results revealed that YOLOv9 achieved the highest inference speed of 332 ms/image at 3.00 Frames Per Second (FPS), while ResNet50 demonstrated superior accuracy in gender classification as a two-stage detection model, albeit with slower performance. To resolve this trade-off, ResNet50 was modified for both speed and accuracy, and then structurally embedded into the YOLOv9 framework. Specifically, the Cross Stage Partial Network (CSPNet) and Efficient Layer Aggregation Network (ELAN) layers of YOLOv9 were replaced with modified ResNet50 feature extractors, while the Global Local Attention Network (GLAN) layer was retained to preserve effective feature fusion. This integration significantly improved both facial and object detection performance. The proposed hybrid model outperformed individual models, achieving a peak mean Average Precision (mAP) of 97.2%, with 97% precision, 93.4% recall, and an inference speed of 103.89 ms/image (9.62 FPS). These results demonstrate that the proposed model effectively balances accuracy and speed, making it highly suitable for real-time applications such as smart surveillance and security systems.

Keywords-deep learning; face detection; gender classification; You Only Look Once (YOLO); ResNet50

I. INTRODUCTION

Over the past two decades, Artificial Intelligence (AI) has achieved remarkable progress, with deep learning emerging as the driving force behind advances in computer vision [1, 2]. A persistent challenge, however, is balancing inference speed and accuracy in real-time detection tasks.

Single-stage detectors such as You Only Look Once (YOLO) are widely adopted for their ability to perform real-time detection, but they often compromise precision [3, 4]. Even the most recent YOLOv9 achieves higher inference speed through optimized processing, though this improvement may come at the expense of accuracy [5, 6]. Moreover, real-world conditions such as variable lighting, diverse facial expressions, and background clutter further complicate detection tasks, reducing overall reliability [7-9].

Several approaches have been proposed in the literature to overcome these challenges. For instance, the Single Shot MultiBox Detector (SSMD) achieves a trade-off between speed and accuracy by performing classification and localization within the same network [10-12]. Two-stage detectors, including Region-based Convolutional Neural Network (R-CNN) and its variants such as Fast and Faster R-CNN, deliver higher accuracy by generating region proposals followed by feature extraction and classification, but they incur substantial computational costs [13-16]. Additionally, ResNet50 has been particularly influential due to its residual connections, which address the vanishing gradient problem and enable deeper feature extraction [17]. It has been successfully applied in practical contexts such as face mask detection [18, 19], and gender classification [20, 21], demonstrating strong accuracy but often with slower inference speed.

To improve efficiency, lightweight CNNs such as MobileNetV2 [22, 23], DenseNet121 [24, 25], EfficientNetB0 [26, 27], and MobileNetV3 [28] were introduced. While these models improve inference time, they often struggle to maintain robust accuracy under complex real-world conditions. More recently, researchers have enhanced YOLO-based frameworks with tailored modules; for example, integrating ResNet50 with focal loss to improve small-object detection without major speed loss [29], further underscoring the ongoing efforts to balance accuracy and efficiency.

In this study, we propose a hybrid framework that combines the real-time detection capability of YOLOv9 [30–34] with the classification accuracy of a structurally modified ResNet50. YOLOv9 was selected for its optimized one-stage detection pipeline, while ResNet50 was modified and embedded to strengthen fine-grained feature extraction for gender classification. Unlike prior works that evaluated YOLO, SSD, or CNNs in isolation [35-37], our approach integrates both paradigms into a unified model to address the fundamental speed–accuracy trade-off. Experimental results on a locally collected dataset demonstrated significant improvements in mean Average Precision (mAP), precision, and recall, while maintaining competitive inference speed. These findings confirm the effectiveness of combining YOLOv9 with a modified ResNet50, making the proposed framework well-suited for real-time applications such as smart surveillance, gender-aware analytics, and security monitoring [38, 39]

II. DATA COLLECTION

The dataset for this study was collected locally from students at the University of Babylon using a smartphone camera, with deliberate consideration of realistic imaging conditions such as variable lighting and diverse facial angles. The dataset contains two classes (male and female) spanning different age groups, and also incorporates partial occlusions to simulate real-world detection challenges. No public or external datasets were used. Explicit informed consent was obtained from all individuals appearing in the images.

III. METHODOLOGY

A. Data Processing

The raw dataset used in this study comprised 140 high-definition facial images, including 72 male and 68 female subjects. To address the limited dataset size and improve model robustness, data augmentation was performed using the Roboflow platform. To simulate real-world variability, advanced augmentation techniques were applied, including horizontal and vertical flipping, rotational adjustments (±15°), random cropping, color shifts (±25%), brightness variations (±25%), and Gaussian blurring (kernel size ≤ 2.5 pixels). These transformations were carefully selected to mimic natural variations in lighting, pose, and image quality, thereby enhancing model generalization. Also, these transformations expanded the dataset from 140 to 980 images while preserving the original male-to-female ratio. The augmented dataset was randomly split into 784 images (80%) for training and 196 images (20%) for validation, ensuring balanced gender representation across subsets. Additionally, an independent test set of 361 images (191 male and 170 female) was reserved exclusively for final evaluation, providing an unbiased benchmark of real-world model performance.

Prior to training, all images were resized to 640 × 640 pixels to match the YOLOv9 input layer, converted to grayscale to emphasize structural facial information, and normalized to the [0,1] range for stable gradient flow. The resolution of 640 × 640 was selected because it represents the default input dimension of YOLOv9, ensuring compatibility with its detection pipeline and preserving the balance between accuracy and computational efficiency.

B. Modified ResNet50 Model

ResNet50 was adopted as the baseline due to its proven feature extraction capability through residual connections. The architecture employed is presented in Figure 1.

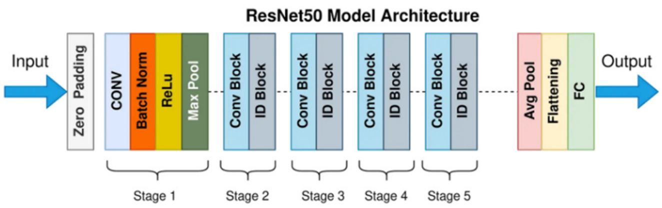

Fig. 1. ResNet50 architecture model.

To tailor ResNet50 for binary gender classification and improve efficiency, several modifications were introduced:

- Activation function: The Rectified Linear Unit (ReLU) activation function was emphasized within the internal convolutional layers to improve gradient stability, accelerate convergence, and mitigate the vanishing gradient problem [40, 41]:

$$\mathrm{ReLU}(x) = \max(0, x) \quad (1)$$

- Global Average Pooling (GAP): A GAP layer was inserted after the convolutional stages to reduce the spatial dimensions of the extracted feature maps. GAP also minimizes parameter count and memory usage while improving generalization by averaging across the feature maps [42, 43], based on (2):

$$F_{\text{pooled}} = \frac{1}{H \times W} \sum_{i=1}^{H} \cdot \sum_{j=1}^{W} \cdot F_{i,j} \quad (2)$$

where F_{pooled} is the value of the features at the position, H and W represents the height and width of the pooling window, and $F_{i,j}$ is the input feature within the pooling window.

- Fully Connected (FC) Layer: The original 1000-class ImageNet FC layer was replaced with a Dense(1) layer followed by a Sigmoid activation, aligning the network with the binary classification task and enabling interpretable probability outputs.
- Transfer learning: The weights of the core convolutional layers were frozen to retain pretrained knowledge while fine-tuning only the newly added layers. This strategy accelerated convergence, preserved the integrity of the learned low-level features, and improved adaptability to the gender classification task.

Through these modifications, the ResNet50 backbone was transformed into a lightweight yet powerful feature extractor with the modified architecture shown in Figure 2.

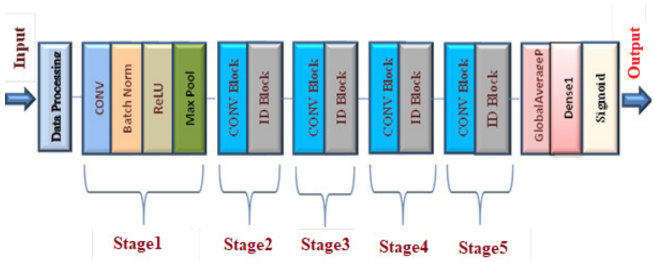

Fig. 2. Proposed modified ResNet50 architecture model.

C. *Proposed Model*

The proposed architecture extends the original YOLOv9 framework by integrating a modified ResNet50 backbone, an enhanced Bidirectional Feature Pyramid Network (BiFPN) + Global-Local Attention Network (GLAN) neck, and a dual-task head for simultaneous face detection and gender classification. The complete workflow of the model is illustrated in Figure 3.

The default Cross-Stage Partial Network (CSPNet) and Efficient Layer Aggregation Network (ELAN) modules in YOLOv9 [44] were replaced with ResNet50, which extracts hierarchical features using residual connections. Input images (640 × 640) were internally resized to 224 × 224 for ResNet50. To harmonize the feature dimensions across scales (P3, P4, P5), Conv2D (1×1) layers are introduced, normalizing the number of channels to 256. This step reduces computational cost without affecting spatial information and ensures compatibility with the subsequent BiFPN stage. Furthermore, a GAP layer is applied after the deepest stage (P5) to minimize the number of parameters and improve generalization.

The default YOLOv9 neck combines FPN and Path Aggregation Network (PANet) for multi-scale feature fusion. However, this configuration suffers from redundancy and lacks adaptive weighting. In the proposed model, the BiFPN replaces the conventional neck, allowing efficient bidirectional flow of information between scales. To further refine multi-scale fusion, weighted feature fusion is introduced, assigning learnable weights to input features, thereby prioritizing the most informative signals. Additionally, a GLAN is embedded after BiFPN to emphasize salient facial regions while suppressing irrelevant background features. BiFPN and GLAN modules reconstructed hierarchical multi-scale feature maps (P3, P4, P5) with spatial resolutions of 80 × 80, 40 × 40, and 20 × 20, ensuring robust multi-scale detection. A final Conv2D (1×1) layer follows GLAN, ensuring consistent feature dimensionality before passing features to the detection head.

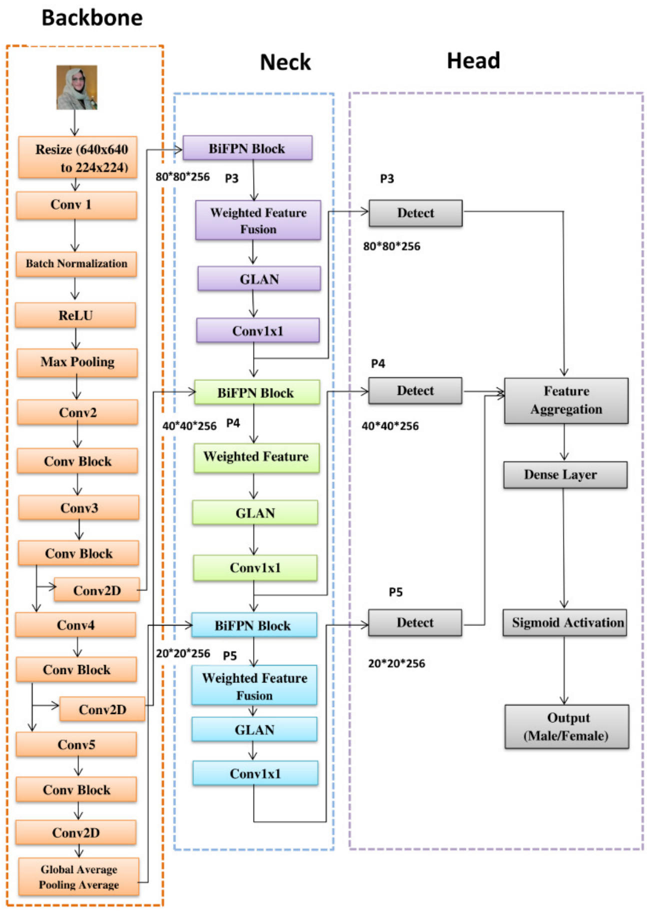

Fig. 3. The design of the proposed improved YOLOv9- modified Resnet50.

The YOLOv9 head is designed for multi-scale detection at P3, P4, and P5, generating bounding boxes, objectness scores, and class probabilities. While this structure performs well for general object detection, it lacks the flexibility required for binary classification tasks such as gender recognition. In the proposed model, the head is extended with a parallel classification branch. Features aggregated from the multi-scale levels are concatenated and passed through a Dense (1) layer with a Sigmoid activation function, producing binary probabilities (male/female). The Sigmoid function is chosen over Softmax due to the binary nature of the task, enabling interpretable probability scores. Furthermore, the auxiliary head present in the original YOLOv9 is removed, as the enhanced backbone and neck already provide sufficiently strong features. This adjustment improves computational efficiency and inference speed without sacrificing accuracy. The hybrid model is mathematically represented as:

$$\hat{y} = \sigma(f_{YOLO}(x, \theta_{YOLO})) \tag{3}$$

where x is the input image, θ_{YOLO} are network weights, f_{YOLO} is the function representing the operations inside the network (such as convolutions, pooling, and other layers), and σ is the Sigmoid activation.

For ResNet50 residual blocks: [45, 46]:

$$y_1 = f_1(x_1, w_1) + x_1 \quad (4)$$

where y_1 is the output of the residual block, f_1 is the function representing the operations within the residual block (such as convolution, batch normalization, and activation), x_1 is the input to the residual block, and w_1 are the weights associated with the operations in the residual block. In a residual block, the input x_1 is added to the output of the function $f_1(x_1, w_1)$. This addition is known as a "skip connection" or "shortcut connection," and it helps to mitigate the vanishing gradient problem by allowing gradients to flow more easily through the network during training.

When the integration of ResNet50 with YOLOv9 replaces the original backbone, it increases the ability to extract complex features more accurately. The modified equation after combining the two algorithms is as follows:

$$\hat{y}_{final} = \sigma\left(f_{YOLOv9_{gender}}(BiFPN((ResNet50(x; \theta_{ResNet}) ; \theta_{BiFPN}); \theta_{YOLOv9/gender})\right) \quad (5)$$

where $\hat{y}_{final}$ is the final predicted output, which is the probability distribution over the possible classes, $\theta_{YOLOv9/gender}$ represents the parameters of the YOLO head for gender classification, $ResNet50(x)$ extracts deep features from the input image, $BiFPN$ performs weighted and precise multi-scale feature fusion across different levels, and θ_{ResNet} represents the parameters of ResNet50 (feature extractor).

D. Training and Testing Setup

The experiments were conducted in Google Colab using an NVIDIA Tesla T4 Graphics Processing Unit (GPU). The model was implemented in PyTorch (v2.0.1) with auxiliary use of TensorFlow and OpenCV for preprocessing. Training was run for 300 epochs with a batch size of 16, while the initial learning rate was set to 0.001 and adjusted dynamically using the ReduceLROnPlateau scheduler, which decreased the rate when the validation loss plateaued, thereby stabilizing training and reducing overfitting risk.

IV. EVALUATION METRICS

The performance of the models in this study was quantified using precision, recall, F1-score, and computational time [47]. These metrics were calculated as follows:

$$\text{Precision} = \frac{TP}{TP+FP} \cdot 100\% \quad (6)$$

$$\text{Recall} = \frac{TP}{TP+FN} \cdot 100\% \quad (7)$$

$$\text{F1} - \text{score} = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (8)$$

where TP denotes True Positives, FP False Positives, TN True Negatives, and FN False Negatives. Additionally, the mAP metric was also employed.

V. RESULTS AND DISCUSSION

To evaluate the performance of the proposed model, a comprehensive comparative analysis of various YOLO versions and classification models (VGG16, DenseNet121, MobileNetV2, and EfficientNetB0) was conducted to determine the most suitable model for object detection. The results of the comparative analysis are presented in Table I.

While YOLOv3 demonstrated effective detection capabilities, its accuracy and speed were suboptimal. YOLOv4 and YOLOv5 improved upon these results but still did not meet the required precision-efficiency trade-off. YOLOv7 provided stronger detection capabilities yet lacked sufficient balance between accuracy and computational time. YOLOv8 introduced substantial improvements; however, YOLOv9 ultimately delivered the best compromise between accuracy and speed, making it the preferred detection model in this study.

TABLE I. PERFORMANCE COMPARISON AMONG DIFFERENT MODELS

Model	mAP (%)	Precision (%)	Recall (%)	F1-score (%)	Inference Time per Image (ms)	Frames Per Second (FPS)	Training time (h:mm:ss)
YOLO v3	83.0	84.3	82.4	83.32	759	1.31	4:00:00
YOLO v4	89.6	85.1	86.6	85.84	648	1.54	3:25:00
YOLOv5	84.1	87.7	83.4	85.49	604	1.65	3:11:00
YOLOv7	94.4	87.9	85.3	86.58	427	2.34	2:15:00
YOLOv8	96.2	88.7	95.2	91.82	379	2.63	2:00:00
YOLOv9	97	86.8	86.1	86.54	332	3.00	1:45:00
ResNet50	94.30	93.60	92	92.79	446.33	2.24	9:18:41
Modified ResNet50	95.60	95	93	93.98	207.73	4.81	6:10:22
VGG16	90	91	92	93.98	618.69	1.62	6:30:00
DenseNet121	91	92	93.80	92.89	172.94	5.78	10:3:00
MobileNetV2	91.80	92.11	92.60	92.35	57.02	17.54	4:30:00
EfficientNetB0	93	93.10	94	93.69	82.91	12.06	5:00:00
YOLOv9-Modified ResNet50 Integrated Model	97.20	97	93.40	95.16	103.89	9.62	3:05:33

Furthermore, classification models were evaluated in two stages to identify the most efficient architecture. VGG16, despite its simplicity, lacked computational efficiency, while DenseNet121 achieved higher accuracy but was computationally expensive. In contrast, MobileNetV2 offered rapid inference but insufficient classification performance. Lastly, EfficientNetB0 provided a reasonable accuracy-speed trade-off but was surpassed by ResNet50, which delivered the highest classification accuracy. Consequently, ResNet50 was selected as the baseline classification model.

A. Performance of Models

Based on Figure 4, YOLOv3 and YOLOv4 showed unstable convergence, with significant oscillations during early epochs and lower final mAP values. YOLOv5 converged more smoothly but became inconsistent in later training stages. In addition, YOLOv7 initially performed well but exhibited sharp degradation, indicating overfitting, while YOLOv8 achieved high accuracy with faster convergence, although minor oscillations were still present. In contrast, YOLOv9 demonstrated rapid and stable convergence, consistently maintaining the highest mAP values across all epochs.

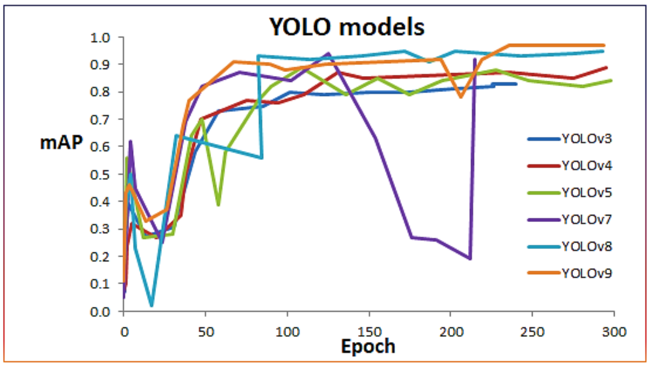

Fig. 4. YOLO models mAP accuracy for gender detection over epochs.

Figure 5 illustrates the training progression of the CNN classifiers over 300 epochs. All models reached stable mAP values after ~50 epochs, but their final performance varied. VGG16 consistently lagged despite gradual improvement. DenseNet121 and EfficientNetB0 delivered strong and stable accuracy but required higher computational resources. MobileNetV2 converged quickly and maintained stability, yet its peak performance was lower than that of ResNet-based models. Both ResNet50 and its modified version achieved superior accuracy, with the modified ResNet50 providing greater stability and the highest mAP throughout training. These results align with Table I, confirming the modified ResNet50 as the most reliable CNN classifier.

Figure 6 compares YOLOv9, Modified ResNet50, and the integrated YOLOv9-Modified ResNet50 model. YOLOv9 exhibited fast early convergence but plateaued at a lower mAP compared to the integrated model. Modified ResNet50 reached high accuracy quickly but showed mild fluctuations. In contrast, the integrated model demonstrated steady and consistent improvement, achieving the best overall performance.

The integrated YOLOv9-Modified ResNet50 model achieved 97.20% mAP and 97.0% precision, surpassing standalone YOLOv9, YOLOv8, and traditional CNNs like ResNet50 and VGG16. Its inference time was 103.89 ms/image (9.62 FPS), slower than MobileNetV2 (57.02 ms/image, 17.54 FPS) but offering much higher accuracy (97.20% vs. 91.80%).

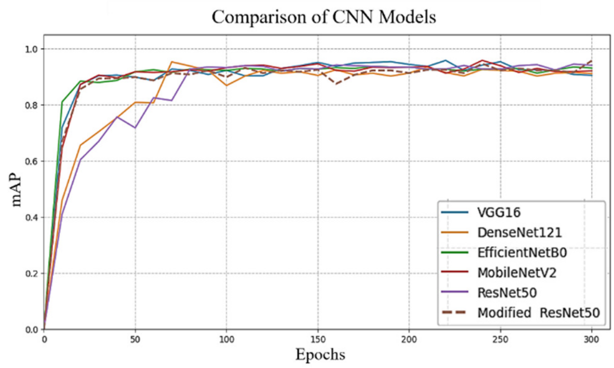

Fig. 5. CNN models mAP accuracy for gender detection over epochs.

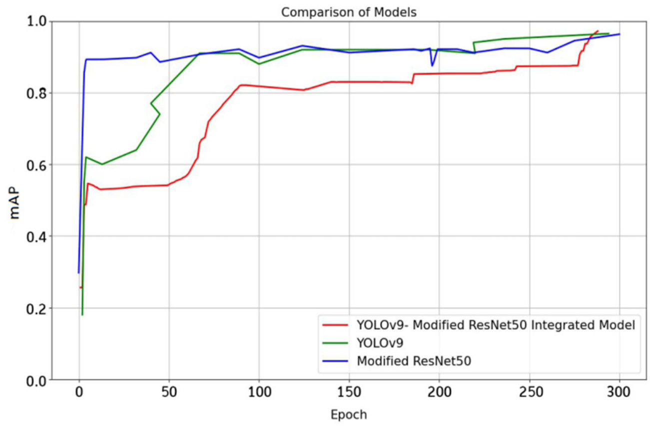

Fig. 6. Comparison (mAP) between modified ResNet50, YOLOv9, and integrated YOLOv9 ResNet50.

B. Loss Object Curve Analysis for Models

As shown in Figure 7, YOLOv9 exhibited a distinct advantage in minimizing object loss over earlier YOLO versions. It achieved a rapid reduction in loss during the initial training stages, stabilizing after ~100 epochs and maintaining consistent performance beyond 150 epochs. This improvement is largely attributed to architectural refinements, such as optimized loss functions and enhanced attention mechanisms, which boost overlap detection and small object detection, the two primary challenges in object detection. While YOLOv8 achieved stable convergence with low loss, it remained less effective than YOLOv9 in advanced training stages. Older models (YOLOv3 and YOLOv4) exhibited substantially higher losses, limiting their effectiveness in complex detection tasks, while YOLOv7 achieved moderate loss reduction but at the cost of training stability.

Among classifiers, the Modified ResNet50 demonstrated the most efficient loss minimization, as illustrated in Figure 8. It began with an initial loss of 1.14, dropping sharply to 0.40 in early epochs and gradually converging to 0.12 in later stages. This behavior reflects both fast convergence and strong stability, confirming its effectiveness in complex classification tasks. DenseNet121 also performed well, decreasing from 1.10 to 0.15, though with reduced stability in later epochs. ResNet50 achieved a final loss of 0.14, while EfficientNetB0 reached 0.13, both competitive but slightly less optimized. In contrast, VGG16 (0.21) and MobileNetV2 (0.11) converged more slowly and less consistently, highlighting their lower efficiency. Overall, the Modified ResNet50 emerged as the most stable and accurate classifier. In terms of inference time, MobileNetV2 (57.02 ms, 17.54 FPS) remained the fastest due to its lightweight design, while EfficientNetB0 (82.91 ms, 12.06 FPS) offered a balanced trade-off. Modified ResNet50 (207.73 ms, 4.81 FPS) enhanced accuracy but was slower than MobileNetV2, and DenseNet121 (172.94 ms, 5.78 FPS), showing moderate efficiency. VGG16 (618.69 ms, 1.62 FPS) was the least efficient due to its outdated architecture.

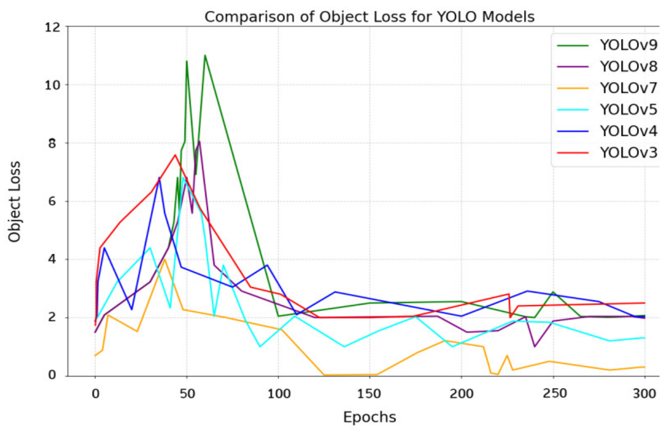

Fig. 7. Comparison of object loss for YOLO models.

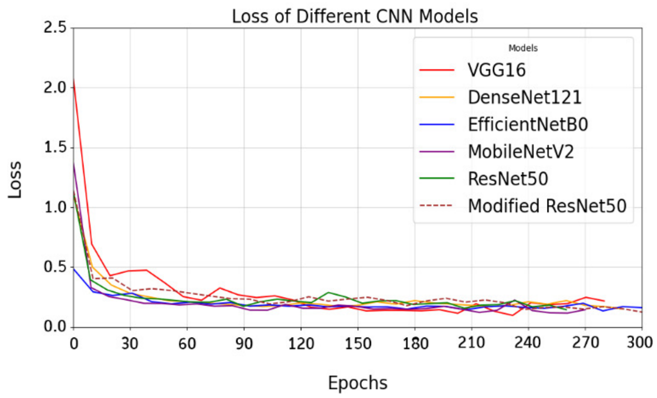

Fig. 8. Comparison of loss for CNN models.

As shown in Figure 9, the integrated YOLOv9-Modified ResNet50 model demonstrated superior performance, achieving rapid loss minimization and greater stability compared to either model alone. While YOLOv9 suffered from fluctuations in early training and Modified ResNet50 converged more steadily, the integrated model combined their strengths, reducing loss quickly and maintaining stability throughout training.

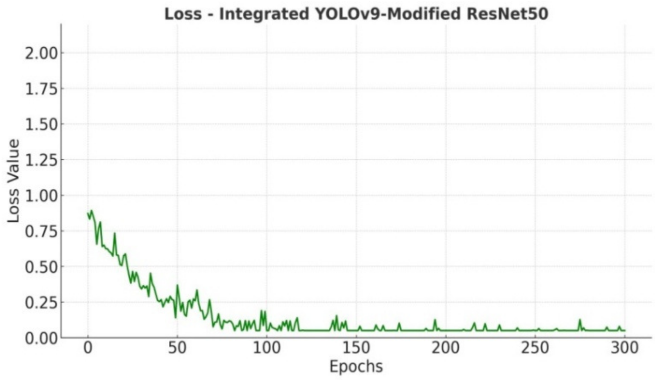

Fig. 9. Loss integrated YOLOv9- modified ResNet50.

C. Confusion Matrix and Error Analysis for Models

The error analysis of YOLO models, depicted in Table II, confirmed YOLOv9's reliability, balancing precision and recall effectively. Although YOLOv8 recorded the lowest total error count of 29, YOLOv9 achieved strong stability with 23 FP and 25 FN, totaling 48 errors. Despite this slightly higher count, YOLOv9 exhibited better consistency and robustness than its predecessors. YOLOv3, with the highest error rate of 59, demonstrated the lowest reliability.

TABLE II. CONFUSION MATRIX ANALYSIS FOR YOLO MODELS

Model	TP	FN	FP	TN
YOLOv3	152	32	27	150
YOLOv4	158	26	23	154
YOLOv5	156	30	21	154
YOLOv7	158	26	21	156
YOLOv8	167	8	21	165
YOLOv9	157	25	23	156

For CNN-based classifiers, confusion matrix analysis (Table III) highlights the Modified ResNet50 as the most accurate model. It correctly classified 173 True Negative (TN) and 168 True Positive (TP) cases, while recording the lowest misclassification rates (FP = 7, FN = 13). EfficientNetB0 and DenseNet121 also delivered strong results with minimal errors, while ResNet50 and MobileNetV2 performed well but with slightly higher error counts. VGG16 exhibited the weakest performance, consistent with its higher loss and slower convergence. The integrated YOLOv9-Modified ResNet50 model achieved the best overall classification performance, with 328 TP, 2 TN, 9 FP, and 22 FN (Figure 10).

Further analysis revealed that most errors occurred in the male class, where some male images were incorrectly classified as female. This slightly reduced the recall rate for males compared to females. The misclassifications were consistent across validation folds and were primarily due to ambiguous facial features or challenging image conditions (e.g., lighting, pose). Because the model's backbone relies on ResNet50-based feature extraction, such edge cases increased the misclassification likelihood. Importantly, these errors do not indicate systematic bias but rather sensitivity to specific visual ambiguities, as shown in Figure 11.

TABLE III. CONFUSION MATRIX ANALYSIS FOR CNN MODELS

Model	TN	FP	FN	TP
ResNet50	170	11	14	166
Modified ResNet50	173	7	13	168
VGG16	162	18	14	167
DenseNet121	165	16	11	169
MobileNetV2	166	15	13	167
EfficientNetB0	168	13	11	169

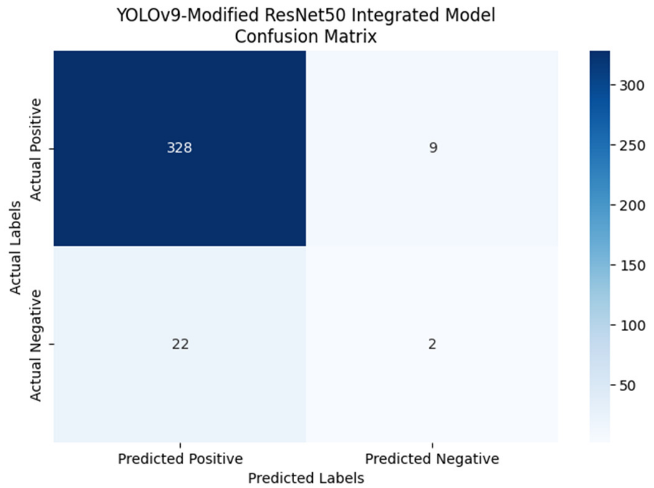

Fig. 10. Confusion matrix of the integrated YOLOv9-Modified ResNet50.

Fig. 11. Examples of male misclassification cases.

D. Statistical Significance Testing

A one-way Analysis of Variance (ANOVA) was performed using Python's SciPy library to assess whether the performance differences among YOLO models (YOLOv3-YOLOv9) were statistically significant. Five independent trials were conducted for each model, with mAP values recorded and summarized in Table IV. The results revealed a significant effect of model type on detection performance (F-value = 8993.63, p-value < 0.05), confirming that the improvements in YOLOv8 and YOLOv9 are not due to random variation but to genuine architectural advances. These findings highlight the impact of optimized backbones, attention mechanisms, and loss functions in enhancing detection accuracy.

A similar ANOVA test was applied to the six CNN-based classifiers to evaluate their statistical differences in gender classification. Results, shown in Table V, confirmed significant variation across models (p-value < 0.05). The Modified ResNet50 achieved the highest accuracy, while VGG16 had the lowest. EfficientNetB0 offered a balanced compromise between accuracy and computational cost, underscoring that model choice should depend on the target application's trade-off between inference speed and accuracy.

A further comparison of parameter sizes is shown in Table VI. The proposed integrated model demonstrates a substantial reduction in parameters compared to the original YOLOv9, owing to the use of a lightweight, modified ResNet50 backbone. By removing the FC layer and employing GAP, computational complexity is reduced while accuracy is preserved or even enhanced. This constitutes a key contribution of the proposed architecture, improving efficiency without compromising detection strength.

TABLE IV. MAP VALUES ACROSS FIVE TRIALS FOR YOLO MODELS

Model	mAP (%)				
Epoch	1	2	3	4	5
YOLOv3	83	82.8	83.1	82.9	83
YOLOv4	89.2	89.8	89.1	88.9	89.6
YOLOv5	84.4	84.1	83.8	84.3	84.1
YOLOv7	93.9	94.2	94.1	94.5	94.4
YOLOv8	96.4	96.1	95.8	96.3	96.2
YOLOv9	97.3	96.9	97.1	97.2	97

TABLE V. MAP VALUES ACROSS FIVE TRIALS FOR CNN MODELS

Model	mAP (%)				
Epoch	1	2	3	4	5
ResNet50	94.6	94.1	94.2	94.4	94.3
Modified ResNet50	95.2	95.4	95.5	95.7	95.6
VGG16	90.9	91.2	91.1	91.5	90.4
DenseNet121	91.2	91.1	91.3	91.5	91
MobileNetV2	91	91.6	91.7	91.9	91.8
EfficientNetB0	93.2	92.9	93.1	93.3	93

TABLE VI. COMPARISON OF MODEL PARAMETERS

Model	Number of Parameters (in Millions)	Notes
Original ResNet50	25.6 M	Includes final FC layer
Modified ResNet50	23.0 M	Final FC layer removed; uses GAP only
Original YOLOv9	64.9 M	Uses CSP + ELAN as the backbone and PANet in the neck
YOLOv9+ Modified ResNet50	27 M	CSP/ELAN replaced with modified ResNet50; PANet replaced with BiFPN + GLAN

E. Feature-Based Analysis (Canny Edge)

To further examine feature extraction, Canny edge detection was implemented in Python using OpenCV. Images were converted to grayscale, denoised via a Gaussian filter, and edges were extracted with the Canny function. The integrated YOLOv9-Modified ResNet50 model successfully leveraged key facial features such as the eyes, eyebrows, mouth, and beard to achieve strong classification performance. However, external factors, including glasses, headscarves, and lighting variations, occasionally reduced accuracy by concealing or distorting critical visual cues.

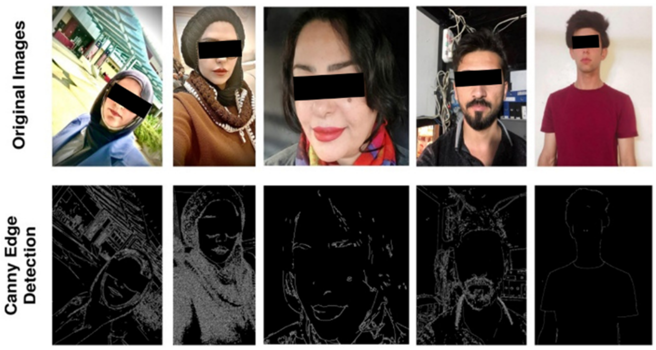

Fig. 12. Canny edge-based feature analysis in YOLOv9-Modified ResNet50 for gender classification.

This analysis provides valuable insight into model behavior, emphasizing the role of preprocessing in enhancing classification robustness. By optimizing training data quality and preprocessing strategies, external variability can be mitigated, enabling more consistent and precise predictions. These results also highlight the potential for integrating additional preprocessing pipelines to improve feature extraction, as shown in Figure 12.

F. Practical Applications and Future Advancements

The proposed YOLOv9-Modified ResNet50 model demonstrated strong potential in real-time object detection and classification, making it suitable for applications in healthcare, security, and autonomous systems for navigation, obstacle detection, and real-time decision-making. Future advancements will focus on enhancing computational efficiency, improving adaptability to diverse environments, and integrating advanced deep learning techniques to refine its performance further.

VI. CONCLUSION

This study presented a comparative evaluation of several deep learning-based detection models, focusing on the distinction between one-stage and two-stage detection algorithms. A hybrid model was proposed, integrating YOLOv9 with a modified lightweight version of ResNet50, aiming to strike an optimal balance between classification accuracy and inference speed. The proposed model demonstrated high effectiveness, achieving 328 True Positive (TP) cases, 9 False Positives (FP), 22 False Negatives (FN), and 14 True Negatives (TN). Evaluation metrics confirmed its superiority with a mean Average Precision (mAP) of 97.20%, precision of 97%, recall of 93.40%, and an F1-score of 95.16%. The key contribution of this work lies in the structural integration of the modified ResNet50 into the YOLOv9 framework, enhancing performance while reducing computational overhead. Unlike previous studies that relied on standalone or conventional models, the proposed hybrid approach offers a robust solution suitable for real-world environments requiring fast response times and high detection accuracy. The results confirm that the proposed model is a promising and scalable solution for real-time computer vision applications, including smart surveillance systems, gender-aware analytics, and autonomous systems. Future enhancements may include incorporating attention mechanisms and optimizing inference speed.

REFERENCES

[1] N. Manakitsa, G. S. Maraslidis, L. Moysis, and G. F. Fragulis, "A Review of Machine Learning and Deep Learning for Object Detection, Semantic Segmentation, and Human Action Recognition in Machine and Robotic Vision," *Technologies*, vol. 12, no. 2, Jan. 2024, Art. no. 15, https://doi.org/10.3390/technologies12020015.

[2] M. Karahan, F. Lacinkaya, K. Erdonmez, E. D. Eminağaoğlu, and C. Kasnakoğlu, "Age and Gender Classification from Facial Features and Object Detection with Machine Learning," *Journal of Fuzzy Extension and Applications*, vol. 3, no. 3, pp. 219-230, Apr. 2022, https://doi.org/10.22105/jfea.2022.328472.1201.

[3] P. T. Anh, N. K. Diep, and N. V Trong, "Convolutional neural networks for image object recognition and classification with large-scale and complex data," *Science & Technology Development Journal-Engineering and Technology*, vol. 6, no. SI8, pp. 10–18, Dec. 2024.

[4] M. A. B. Zuraimi and F. H. K. Zaman, "Vehicle Detection and Tracking using YOLO and DeepSORT," in *2021 IEEE 11th IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE)*, Penang, Malaysia, Apr. 2021, pp. 23–29, https://doi.org/10.1109/ISCAIE51753.2021.9431784.

[5] P. Mahto, P. Garg, P. Seth, and J. Panda, "Refining Yolov4 for Vehicle Detection", *International Journal of Advanced Research in Engineering and Technology (IJARET)*, vol. 11, no. 5, pp. 409-419, Jul. 2020.

[6] F. E. Ayo, A. M. Mustapha, J. A. Braimah, and D. A. Aina, "Geometric Analysis and YOLO Algorithm for Automatic Face Detection System in a Security Setting," *Journal of Physics: Conference Series*, vol. 2199, no. 1, Feb. 2022, Art. no. 012010, https://doi.org/10.1088/1742-6596/2199/1/012010.

[7] E. K. Varnima and C. Ramachandran, "Real-time Gender Identification from Face Images using you only look once (yolo)," in *2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184)*, Tirunelveli, India, Jun. 2020, pp. 1074–1077, https://doi.org/10.1109/ICOEI48184.2020.9142989.

[8] W. Ali, W. Tian, S. U. Din, D. Iradukunda, and A. A. Khan, "Classical and modern face recognition approaches: a complete review," *Multimedia Tools and Applications*, vol. 80, no. 3, pp. 4825–4880, Jan. 2021, https://doi.org/10.1007/s11042-020-09850-1.

[9] S. Umer, B. C. Dhara, and B. Chanda, "Face recognition using fusion of feature learning techniques," *Measurement*, vol. 146, pp. 43–54, Nov. 2019, https://doi.org/10.1016/j.measurement.2019.06.008.

[10] W. Chen, H. Huang, S. Peng, C. Zhou, and C. Zhang, "YOLO-face: a real-time face detector," *The Visual Computer*, vol. 37, no. 4, pp. 805–813, Apr. 2021, https://doi.org/10.1007/s00371-020-01831-7.

[11] A. N. Kadhum and A. N. Kadhum, "Literature Survey on YOLO Models for Face Recognition in Covid-19 Pandemic," *Journal of Image Processing and Intelligent Remote Sensing*, no. 34, pp. 27–35, Jul. 2023, https://doi.org/10.55529/jipirs.34.27.35.

[12] S. Wang, "Design of smart community access control system based on SSD and OneNET cloud platform," in *3rd International Conference on Internet of Things and Smart City (IoTSC 2023)*, Chongqing, China, Jun. 2023, Art. no. 104, https://doi.org/10.1117/12.2684043.

[13] W. Chen, Y. Qiao, and Y. Li, "Inception-SSD: An improved single shot detector for vehicle detection," *Journal of Ambient Intelligence and Humanized Computing*, vol. 13, no. 11, pp. 5047–5053, Nov. 2022, https://doi.org/10.1007/s12652-020-02085-w.

[14] A. Dhillon and G. K. Verma, "Convolutional neural network: a review of models, methodologies and applications to object detection," *Progress in Artificial Intelligence*, vol. 9, no. 2, pp. 85–112, Jun. 2020, https://doi.org/10.1007/s13748-019-00203-0.

[15] R. Kaur and S. Singh, "A comprehensive review of object detection with deep learning," *Digital Signal Processing*, vol. 132, Jan. 2023, Art. no. 103812, https://doi.org/10.1016/j.dsp.2022.103812.

[16] A. Mustafa and K. Meehan, "Gender Classification and Age Prediction using CNN and ResNet in Real-Time," in *2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI)*, Sakheer, Bahrain, Oct. 2020, pp. 1–6, https://doi.org/10.1109/ICDABI51230.2020.9325696.

[17] H. Jiang and E. Learned-Miller, "Face Detection with the Faster R-CNN," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, Washington, DC, DC, USA, May 2017, pp. 650–657, https://doi.org/10.1109/FG.2017.82.

[18] M. Othmani, "A vehicle detection and tracking method for traffic video based on faster R-CNN," *Multimedia Tools and Applications*, vol. 81, no. 20, pp. 28347–28365, Aug. 2022, https://doi.org/10.1007/s11042-022-12715-4.

[19] S. D. Meena, C. S. Siri, P. S. Lakshmi, N. S. Doondı, and J. Sheela, "Real time DNN-based Face Mask Detection System using MobileNetV2 and ResNet50," in *2023 International Conference on Inventive Computation Technologies (ICICT)*, Lalitpur, Nepal, Apr. 2023, pp. 1007–1015, https://doi.org/10.1109/ICICT57646.2023.10133992.

[20] X. Zhao, "Research and application of deep learning in image recognition," *Journal of Physics: Conference Series*, vol. 2425, no. 1, Feb. 2023, Art. no. 012047, https://doi.org/10.1088/1742-6596/2425/1/012047.

[21] A. Ahmed and F. S. Alghareb, "A Hybrid ROI Extraction Approach for Mask and Unmask Facial Recognition System using Light-CNN," *International Journal of Computing and Digital Systems*, vol. 15, no. 1, pp. 1223–1232, Sep. 2024, https://doi.org/10.12785/ijcds/160190.

[22] C. Lin, X. Hu, Y. Zhan, and X. Hao, "MobileNetV2 with Spatial Attention module for traffic congestion recognition in surveillance images," *Expert Systems with Applications*, vol. 255, Dec. 2024, Art. no. 124701, https://doi.org/10.1016/j.eswa.2024.124701.

[23] S. Dodia, V. Meshram, J. Kasle, S. Gomase, H. Amrit, and R. Sarse, "Autism Spectrum Disorder (ASD) Detection from Facial Images using MobileNet," in *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)*, Pune, India, Apr. 2024, pp. 1–7, https://doi.org/10.1109/I2CT61223.2024.10543439.

[24] J. Zhang *et al.*, "Hyperspectral Image Classification Based on Dense Pyramidal Convolution and Multi-Feature Fusion," *Remote Sensing*, vol. 15, no. 12, Art. no. 2990, Jun. 2023, https://doi.org/10.3390/rs15122990.

[25] L. Kong and J. Cheng, "Classification and detection of COVID-19 X-Ray images based on DenseNet and VGG16 feature fusion," *Biomedical Signal Processing and Control*, vol. 77, Aug. 2022, Art. no. 103772, https://doi.org/10.1016/j.bspc.2022.103772.

[26] J. X. Mi, J. Feng, and K.-Y. Huang, "Designing efficient convolutional neural network structure: A survey," *Neurocomputing*, vol. 489, pp. 139–156, Jun. 2022, https://doi.org/10.1016/j.neucom.2021.08.158.

[27] Z. Huang, X. Jiang, S. Huang, S. Qin, and S. Yang, "An efficient convolutional neural network-based diagnosis system for citrus fruit diseases," *Frontiers in Genetics*, vol. 14, Aug. 2023, Art. no. 1253934, https://doi.org/10.3389/fgene.2023.1253934.

[28] S. Ennaama, H. Silkan, A. Bentajer, and A. Tahiri, "Enhanced Real-Time Object Detection using YOLOv7 and MobileNetv3," *Engineering, Technology & Applied Science Research*, vol. 15, no. 1, pp. 19181–19187, Feb. 2025, https://doi.org/10.48084/etasr.8777.

[29] M. Shafiq and Z. Gu, "Deep Residual Learning for Image Recognition: A Survey," *Applied Sciences*, vol. 12, no. 18, Sep. 2022, Art. no. 8972, https://doi.org/10.3390/app12188972.

[30] P. Jabraelzadeh, A. Charmin, and M. Ebadpour, "Providing a hybrid method for face detection and gender recognition by a transfer learning and fine-tuning approach in deep convolutional neural networks and the Yolo algorithm," *International Journal of Nonlinear Analysis and Applications*, vol. 14, no. 1, pp. 2373-2381, Jul. 2022, https://doi.org/10.22075/ijnaa.2022.26099.3661.

[31] S. Minaee, P. Luo, Z. Lin, and K. Bowyer, "Going Deeper Into Face Detection: A Survey." arXiv, Apr. 13, 2021, https://doi.org/10.48550/arXiv.2103.14983.

[32] Z. M. Peerun and R. K. Moloo, "Real-time gender and people tracking using YOLO," in *2024 Sixth International Conference on Computational Intelligence and Communication Technologies (CCICT)*, Sonepat, India, Apr. 2024, pp. 448–454, https://doi.org/10.1109/CCICT62777.2024.00079.

[33] V. Viswanatha, R. K. Chandana, and A. C. Ramachandra, "Real Time Object Detection System with YOLO and CNN Models: A Review." arXiv, 2022, https://doi.org/10.48550/ARXIV.2208.00773.

[34] A. Nowrin, S. Afroz, Md. S. Rahman, I. Mahmud, and Y.-Z. Cho, "Comprehensive Review on Facemask Detection Techniques in the Context of Covid-19," *IEEE Access*, vol. 9, pp. 106839–106864, 2021, https://doi.org/10.1109/ACCESS.2021.3100070.

[35] M. Patel and U. Singh, "Age and Gender Recognition using Deep Learning Technique," in *2023 3rd International Conference on Smart Data Intelligence (ICSMDI)*, Trichy, India, Mar. 2023, pp. 238–245, https://doi.org/10.1109/ICSMDI57622.2023.00052.

[36] M. N. A. Aziz, S. Mutalib, and S. Aliman, "Comparison of Face Coverings Detection Methods using Deep Learning," in *2021 2nd International Conference on Artificial Intelligence and Data Sciences (AiDAS)*, IPOH, Malaysia, Sep. 2021, pp. 1–6, https://doi.org/10.1109/AiDAS53897.2021.9574318.

[37] I. Oztel, "Human Detection System using Different Depths of the Resnet-50 in Faster R-CNN," in *2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, Istanbul, Turkey, Oct. 2020, pp. 1–5, https://doi.org/10.1109/ISMSIT50672.2020.9255109.

[38] J. E. Gallagher and E. J. Oughton, "Surveying You Only Look Once (YOLO) Multispectral Object Detection Advancements, Applications, and Challenges," *IEEE Access*, vol. 13, pp. 7366–7395, 2025, https://doi.org/10.1109/ACCESS.2025.3526458.

[39] X. Guo, Y.-D. Zhang, S. Lu, and Z. Lu, "A Survey on Machine Learning in COVID-19 Diagnosis," *Computer Modeling in Engineering & Sciences*, vol. 130, no. 1, pp. 23–71, 2022, https://doi.org/10.32604/cmes.2021.017679.

[40] Y. Feng, M. Gao, and Z. Zhang, "Web Service QoS Classification Based on Optimized Convolutional Neural Network," in *2019 IEEE 14th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*, Dalian, China, Nov. 2019, pp. 584–590, https://doi.org/10.1109/ISKE47853.2019.9170368.

[41] E. Yildirim, "ResNet-based Gender Recognition on Hand Images," *Engineering, Technology & Applied Science Research*, vol. 14, no. 6, pp. 17969–17972, Dec. 2024, https://doi.org/10.48084/etasr.8922.

[42] P. Dey, T. Mahmud, M. S. Chowdhury, M. S. Hossain, and K. Andersson, "Human Age and Gender Prediction from Facial Images Using Deep Learning Methods," *Procedia Computer Science*, vol. 238, pp. 314–321, 2024, https://doi.org/10.1016/j.procs.2024.06.030.

[43] C. Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall, "Activation Functions: Comparison of trends in Practice and Research for Deep Learning." arXiv, 2018, https://doi.org/10.48550/arXiv.1811.03378.

[44] S. Chaudhuri *et al.*, "Infrared Thermography of Turbulence Patterns of Operational Wind Turbine Rotor Blades Supported With High-Resolution Photography: KI-VISIR Dataset," *Wind Energy*, vol. 28, no. 1, Jan. 2025, Art. no. e2958, https://doi.org/10.1002/we.2958.

[45] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection." arXiv, 2015, https://doi.org/10.48550/ARXIV.1506.02640.

[46] S. Mukherjee. "The Annotated ResNet-50: Explaining how ResNet-50 works and why it is so popular," Towards Data Science. [Online]. Available: https://towardsdatascience.com/the-annotated-resnet-50-a6c536034758.

[47] T. Szandała, "Review and Comparison of Commonly Used Activation Functions for Deep Neural Networks," in *Bio-inspired Neurocomputing*, vol. 903, A. K. Bhoi, P. K. Mallick, C.-M. Liu, and V. E. Balas, Eds. Singapore: Springer Singapore, 2021, pp. 203–224.