



OPEN A hybrid object detection approach for visually impaired persons using pigeon-inspired optimization and deep learning models

Abdullah M. Alashjaee¹, Hussah Nasser AlEisa², Abdulbasit A. Darem^{3,4}✉ & Radwa Marzouk⁵

Visually challenged persons include a significant part of the population, and they exist all over the globe. Recently, technology has demonstrated its occurrence in each field, and state-of-the-art devices aid humans in their everyday lives. However, visually impaired people cannot view things around their atmospheres; they can only imagine the roaming surroundings. Furthermore, web-based applications are advanced to certify their protection. Using the application, the consumer can spin the requested task to share her/his position with the family members while threatening confidentiality. Through this application, visually challenged people's family members can follow their actions (acquire snapshots and position) while staying at their residences. A deep learning (DL) technique is trained with manifold images of entities highly related to the VIPs. Training images are amplified and physically interpreted to bring more strength to the trained method. This study proposes a Hybrid Approach to Object Detection for Visually Impaired Persons Using Attention-Driven Deep Learning (HAODVIP-ADL) technique. The major intention of the HAODVIP-ADL technique is to deliver a reliable and precise object detection system that supports the visually impaired person in navigating their surroundings safely and effectively. The presented HAODVIP-ADL method initially utilizes bilateral filtering (BF) for the image pre-processing stage to reduce noise while preserving edges for clarity. For object detection, the HAODVIP-ADL method employs the YOLOv10 framework. In addition, the backbone fusion of feature extraction models such as CapsNet and InceptionV3 is implemented to capture diverse spatial and contextual information. The bi-directional long short-term memory and multi-head attention (MHA-BiLSTM) approach is utilized to classify the object detection process. Finally, the hyperparameter tuning process is performed using the pigeon-inspired optimization (PIO) approach to advance the classification performance of the MHA-BiLSTM approach. The experimental results of the HAODVIP-ADL method are analyzed, and the outcomes are evaluated using the Indoor Objects Detection dataset. The experimental validation of the HAODVIP-ADL method portrayed a superior accuracy value of 99.74% over the existing methods.

Keywords Object detection, Visually impaired persons, Deep learning, Pigeon-inspired optimization, Feature extraction

A human's core feature is usually vision ability. The ability to see things with the eyes is considered a gift and a significant factor in everyday activities. A major problem in several visually impaired people is that they cannot be independent and are imperfect in their vision¹. Visually impaired individuals face difficulties with such activities, and object recognition is an essential factor on which they can depend frequently. They regularly have problems identifying objects and surrounding movements, mainly walking along the street². Most humans have a vision, as witnessed at the age of fifty. Visually impaired people depend on their auditory perception, braille, and somatosensation- primarily sound- to gain information from the environment; they utilize helpful devices like canes to identify obstacles³. Although 28.22% of the worldwide population comprises visually impaired individuals, obtainable facilities aren't generally installed, which leads to social discrimination issues owing

¹Department of Computer Science, College of Science, Northern Border University, Arar, Saudi Arabia. ²Department of Computer Sciences, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. ³Center for Scientific Research and Entrepreneurship, Northern Border University, Arar 73213, Saudi Arabia. ⁴King Salman Center for Disability Research, Riyadh 11614, Saudi Arabia. ⁵Department of Mathematics, Faculty of Science, Cairo University, Giza 12613, Egypt. ✉email: basit.darem@nbu.edu.sa

to the confines of their activities⁴. They cannot independently face unexpected situations outdoors, confining their indoor activities. It is typically a challenge for machines to classify and distinguish multiple things in an image. Within the Computer Vision (CV) domain, object detection relates to identifying and detecting an object present in a video or image⁵. Nevertheless, object detection has been a significant effort in recent years. The key elements of object detection comprise processing, object classification, and feature extraction⁶.

Various models involving feature aggregation, feature coding, feature classification, bottom feature extraction, and object detection have been utilized to create satisfying outcomes between these methods; feature extraction plays a vital role in process recognition and object detection. Object detection is essential in many applications, including but not finite surveillance, vehicle detection, cancer diagnosis, and object identification in outdoor and indoor environments⁷. Artificial Intelligence (AI) is initiating novel methods for individuals with disabilities to admittance to the globe. Object recognition is one of the earlier utilization of AI by individuals who have low vision or are blind in apps like Seeing AI⁸. However, generic AI object recognition is limited to discovering popular objects, like room doors or chairs. Consequently, attention has transferred to manageable object recognizers that can assist users in identifying objects: (1) Without coverage of standard generic classes, for example, white canes; (2) Specified instances of an object⁹. Specialists employed Deep Learning (DL) to directly attain features from image pixels in search of object recognition and detection. This method demonstrated more significant results than classical Deep Neural Network (DNN) methods, especially region-based Convolutional Neural Networks (CNNs). Object detection approaches depend upon DL, and those are given better performances on every approach¹⁰. They are separated into two major groups: (1) One-stage approaches that are performed by the object classification and localization in a unique network, and (2) two-stage techniques that have two scattered systems for classification and localization.

This study proposes a Hybrid Approach to Object Detection for Visually Impaired Persons Using Attention-Driven Deep Learning (HAODVIP-ADL) technique. The major intention of the HAODVIP-ADL technique is to deliver a reliable and precise object detection system that supports the visually impaired person in navigating their surroundings safely and effectively. The presented HAODVIP-ADL method initially utilizes bilateral filtering (BF) for the image pre-processing stage to reduce noise while preserving edges for clarity. For object detection, the HAODVIP-ADL method employs the YOLOv10 framework. In addition, the backbone fusion of feature extraction models such as CapsNet and InceptionV3 is implemented to capture diverse spatial and contextual information. The bi-directional long short-term memory and multi-head attention (MHA-BiLSTM) approach is utilized to classify the object detection process. Finally, the hyperparameter tuning process is performed using the pigeon-inspired optimization (PIO) approach to advance the classification performance of the MHA-BiLSTM approach. The experimental results of the HAODVIP-ADL method are analyzed, and the outcomes are evaluated using the Indoor Objects Detection dataset.

- BF-based pre-processing improves image quality by effectively mitigating noise while preserving crucial edges. This step confirms that the input images are cleaner and more focused, resulting in improved performance in object detection and feature extraction models. It also assists in enhancing overall model accuracy by giving high-quality input.
- The YOLOv10-based object detection enables accurate detection and localization of regions of interest within images. This advanced methodology ensures precise detection, which is significant for subsequent feature extraction and classification tasks. It improves the technique's capability to concentrate on relevant areas, enhancing overall analysis efficiency.
- The model can better comprehend local and global data patterns by incorporating CapsNet for capturing spatial hierarchies and InceptionV3 for deep feature extraction. This integration enables the model to capture complex and intricate features efficiently, improving overall performance by improving feature representation.
- The hybrid approach of BiLSTM and MHA implements the merits of both models for improved temporal and contextual understanding. BiLSTM captures sequential dependencies, while MHA concentrates on significant features across diverse attention heads. This incorporation enhances classification accuracy by effectively handling complex patterns in data.
- Integrating PIO-based tuning presents a novel optimization strategy by replicating natural pigeon behaviour for parameter fine-tuning. This method enables the technique to explore the parameter space more efficiently than conventional optimization techniques. By utilizing PIO, the model attains enhanced performance and accuracy, surpassing the capabilities of traditional methods in fine-tuning and optimization.

Related works

Dang et al.¹¹ examine implementing and developing a DL-based mobile application to help visually impaired and blind individuals in real-world pill identification. Employing the YOLO structure, this application intends to precisely detect and discriminate between several pill types over real-world image processing on mobile devices. This method integrates Text-to-Speech (TTS) to offer immediate auditory feedback, independence for visually impaired users, and improving usability. In¹², an enhanced YOLOv5 method was presented, and the coupled head was substituted with an improved decoupled head. This method is also applied to enhance activation and loss functions. Additionally, the TTS conversion Google library is employed to detect the conversion class of the object into an audio signal to deliver information to visually impaired individuals. In¹³, a new gadget attached to standard eyeglasses exploits advanced DL models to develop real-world banknote recognition and detection. The YOLOv5 method was chosen as the best-performing method. In addition, this research compared the YOLOv5 model with standard object detection methods like Faster_RCNN_Inception_v2 and SSD_MobileNet_v2. Yannawar¹⁴ projects a solution method to support the visually impaired (VI) population. This method aimed at object detection and feature extraction utilizing a Convolutional Neural Network (CNN) from a real-world video. For this, a head-mounted image acquisition gadget might be employed to identify the

objects from the scene ahead, and information on the recognized objects is delivered to VI individuals through the audio modality. Oluyele et al.¹⁵ developed a CNN-based object recognition method incorporated into a mobile robot that functions as a robotic assistant for VI individuals. The robotic assistant was able to move around in a confined environment.

Triyono et al.¹⁶ project a complete collision detection and prevention method. The presented method incorporates cutting-edge technologies, comprising audio production devices, DL, image processing, cloud computing, and IoT. By associating these technologies with the white cane, this method provides an advanced navigation option for the VI, effectively preventing and detecting possible collisions. Shilaskar et al.¹⁷ intend to design a smart walker that assists VI people in enhancing their stability and mobility. The developed method utilizes an association of ultrasonic sensors, computer vision (CV), and real-world audio feedback to raise the flexibility and safety of blind individuals through navigation. The Smart Walker has an onboard camera, which scans the environment and utilizes CV methods to detect complications and evaluate the object's distance. Rocha et al.¹⁸ utilize object detection technology. The defect detection method projected in this research depends on the YOLO structure, a single-stage object detector suitable for automatic inspection tasks. This method employed for the optimization of defect detection depended on 3 major modules: (1) raising the dataset with novel defects, backgrounds, and illumination conditions, (2) presenting data augmentation, and (3) initiating defect classification. Wang et al.¹⁹ propose a DL approach utilizing video surveillance. It integrates spatial feature extraction with U-Net, MobileNetV2 for encoding, and an enhanced LSTM method for temporal feature extraction. Optical flow is used to track individuals and crowds, enhancing surveillance. Mencattini et al.²⁰ present a platform for classifying HEP-2 images and ANA patterns utilizing ResNET101 for feature extraction and SVM and LDA for classification. It comprises feature selection, majority voting, and Grad-CAM for explainability, addressing unbalanced datasets, cross-hardware compatibility, and performance evaluation with a sample quality index. Mohanty et al.²¹ aim to improve brain lesion image classification using advanced DL approaches and the African vulture optimization (AVO) method, focusing on improving accuracy through feature fusion and grid-based techniques.

Zhou et al.²² propose the spatial-spectral cross-attention-driven network (SSCA-DN) network for improved spatial-spectral reconstruction. It incorporates multi-scale feature aggregation (MFA) and a spectral transformer (SpeT) to capture spatial-spectral correlations. The network utilizes supervised and unsupervised subnetworks to refine features and improve reconstruction accuracy. Madarapu et al.²³ introduce the multi-resolution convolutional attention network (MuR-CAN) methodology by utilizing a multi-dilation attention block to capture multi-scale features and enhance DR classification accuracy through integration with an SVM classifier. Cheng et al.²⁴ present the Dynamic Interactive Network with Self-Distillation (DISD-Net) for cross-subject MER to capture intra- and inter-modal interactions. Self-distillation improves modal representations by transferring knowledge through hard and soft labels. Domain adaptation (DA) is integrated to extract subject-invariant multi-modal emotional features in a unified framework. Ravinder and Srinivasan²⁵ propose a DL approach for captioning medical images, incorporating YOLOv4 for feature extraction and a recurrent neural network (RNN) with LSTM and attention for generating accurate, context-aware descriptions of defects in images. Luan et al.²⁶ propose FMambaIR, an image restoration model that integrates frequency and Mamba for improved restoration. The F-Mamba block combines spatial Mamba and Fourier frequency-domain modelling to capture global degradation. A forward feedback network is utilized to refine local details for improved recovery. Jaiswal et al.²⁷ present a novel image captioning method with a dual self-attention encoder-decoder framework, utilizing VGG16 Hybrid Places 1365 for feature extraction and gated recurrent unit (GRU) for word-level language modelling, improving contextual image description. Banu et al.²⁸ explore the role of visual information processing in computational neuroscience and healthcare, highlighting advancements in DL, neural networks, and AI technologies. It accentuates their impact on understanding brain activity, medical diagnostics, and treatment personalization while addressing ethical and data privacy concerns.

Despite improvements in DL-based applications for the VI, threats remain in real-time processing, accuracy, and robustness, mainly when dealing with complex, dynamic environments. Current methods mostly face difficulty with varying lighting conditions, background noise, and object detection accuracy, which can affect performance in real-world scenarios. Furthermore, many existing solutions lack adaptability to diverse subjects or environments, restricting their generalization capabilities. Though promising, incorporating self-distillation and domain adaptation requires additional refinement for improved accuracy. Moreover, a significant gap remains: a lack of cross-domain evaluation and real-time deployment in mobile devices. Finally, while methods such as YOLO and CNN exhibit promise, they often require more computational power, impacting usability in portable solutions for the VI.

Materials and methods

This study presents a HAODVIP-ADL model. The main intention of the model is to deliver a reliable and precise object detection system that supports the visually impaired in navigating their surroundings safely and effectively. The proposed HAODVIP-ADL model involves various stages like image pre-processing, object detection, feature extraction, classification process, and hyperparameter tuning model to accomplish this. The overall working process of the HAODVIP-ADL approach is depicted in Fig. 1.

Image pre-processing using BF

The presented HAODVIP-ADL method initially applies BF for the image pre-processing stage to reduce noise while preserving edges for clarity²⁹. This model is chosen due to its capability to effectively mitigate noise while conserving edges, which is significant for maintaining crucial structural details in the image. Unlike conventional smoothing techniques, BF can distinguish between noise and edges by considering spatial distance and intensity differences, resulting in sharper and more accurate feature extraction. This makes it appropriate

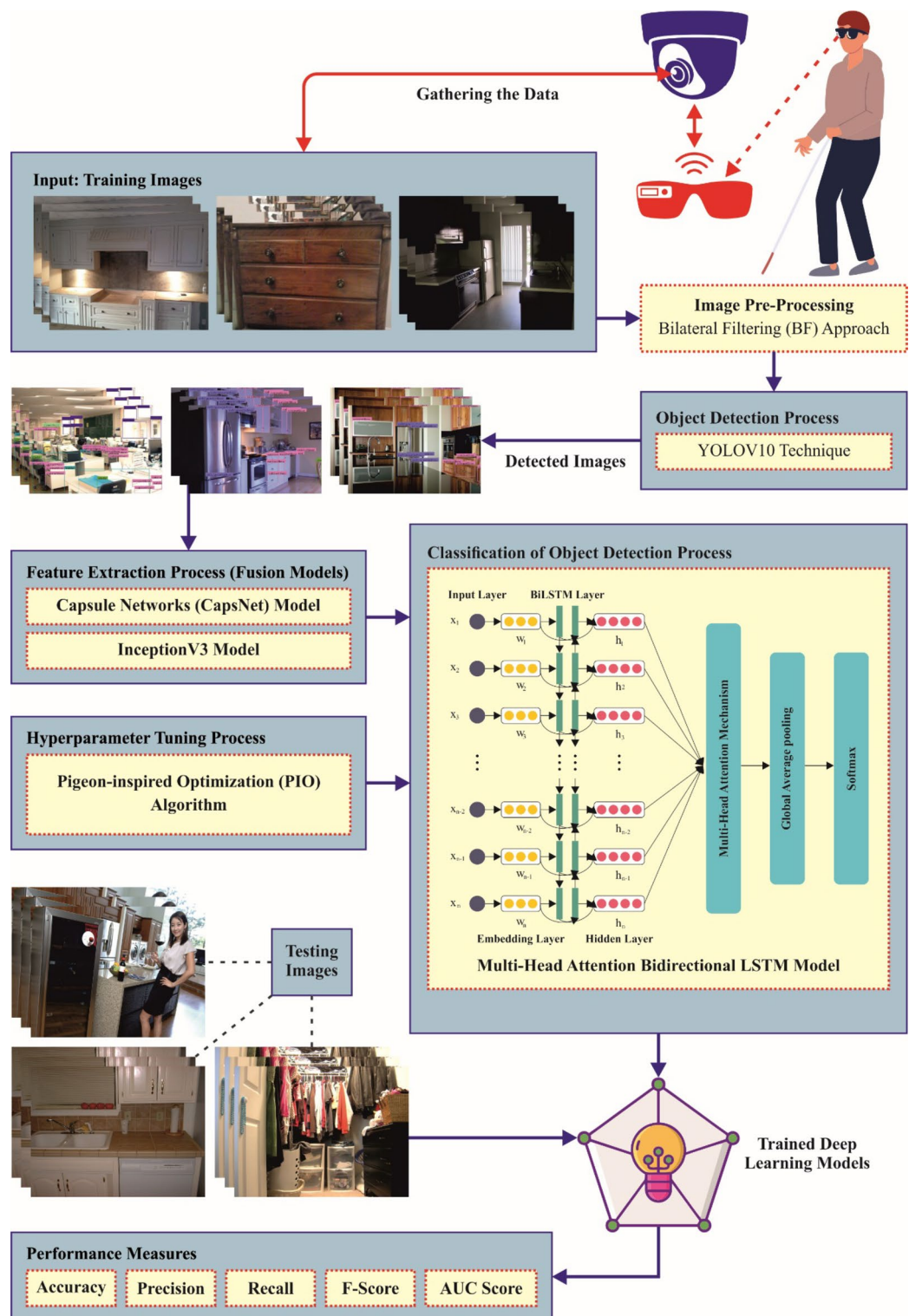


Fig. 1. Overall workflow of the HAODVIP-ADL model.

for applications needing high-quality image inputs, such as object detection and classification. Furthermore, BF's nonlinear nature allows it to adapt to varying image complexities, giving superior performance in diverse conditions related to linear filters like Gaussian blur. Its computational efficiency and adaptability make it an ideal choice over more complex methods like anisotropic diffusion or wavelet-based denoising. Figure 2 specifies the BF framework.

BF is a method employed in image pre-processing to smooth images while maintaining edges, making it perfect for the detection of object tasks. For visually impaired people, it can improve the dissimilarity of objects,

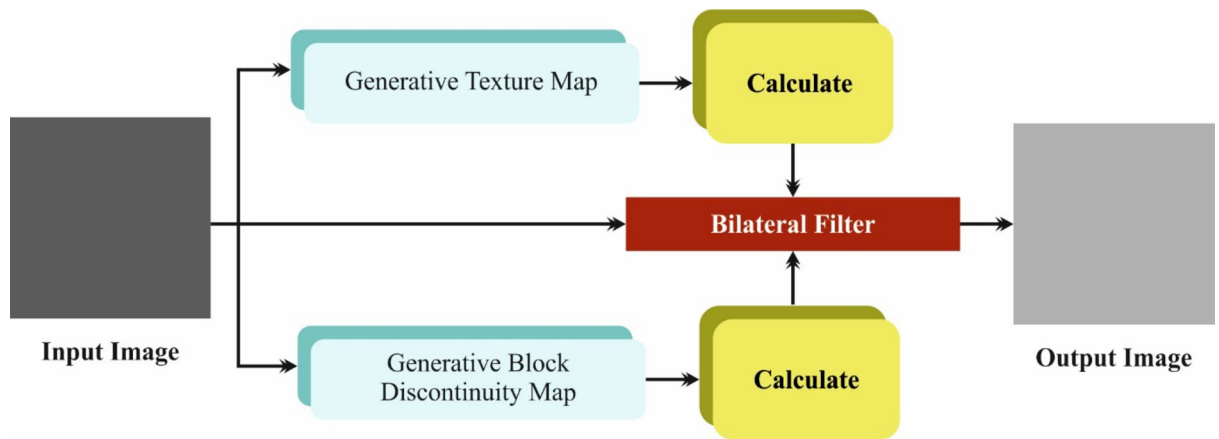


Fig. 2. Structure of BF model.

particularly in challenging visual conditions. By decreasing noise and upholding object limits, bilateral filtering aids enhance the accuracy of object recognition methods, certifying they can recognize related features in the environment. The model functions by considering colour similarity and pixel spatial proximity, which helps differentiate objects from their surroundings. This can be the main reason for more trustworthy navigation help over AI-powered methods for visually challenged people. Combining BF with object detection techniques improves real feedback, vital for situational alertness and obstacle avoidance.

Object detection using YOLOV10 framework

For object detection, the HAODVIP-ADL model utilized the YOLOv10 framework³⁰. This model is chosen due to its superior speed and accuracy, making it ideal for real-time applications. YOLOv10 benefits from an enhanced architecture that improves feature extraction and prediction accuracy related to previous YOLO versions, maintaining high detection precision while processing images at faster rates. Unlike region-based methods, namely Faster R-CNN, which are slower due to multiple stages, YOLOv10 is a single-stage detector with lower latency and better performance in dynamic environments. The capability of the method to detect various objects in diverse scales within a single pass, along with its robustness to small and occluded objects, sets it apart from conventional CNN-based detectors. Moreover, its lightweight architecture makes it appropriate for mobile and embedded systems, where computational resources are limited. Figure 3 demonstrates the YOLOv10 methodology.

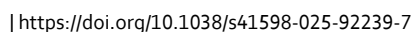
The latest form of the YOLO series during this study is YOLOv10. This form is designed to address the remaining problems of the YOLO series: accuracy, speed, and efficiency. Initially, to enhance inference efficacy, authors presented a substitute for Non-Maximum Suppression (NMS), a computationally costly post-processing method applied in preceding forms of YOLO to deal with the problem of numerous and redundant predictions. Instead, the authors present a constant two-label task in which every object is allocated through a novel bounding box. This is attained by joining the best of either tactic: one-to-one (o2o) task and one-to-many (o2m) task. Both approaches utilize a uniform matching metric:

$$m(\alpha, \beta) = s \cdot p^\alpha \cdot IoU(\hat{b}, b)^\beta \quad (1)$$

Whereas p denotes the classification score, b and \hat{b} are ground truth and prediction bounding boxes correspondingly, and s corresponds with an anchor prior. α and β represent hyperparameters that balance classification and semantic tasks. A metric is computed for all approaches, m_{o2o} and m_{o2m} . The $o2m$ presents a richer supervisory signal, and accordingly, the optimization of $o2o$ is harmonized. Therefore, the o2o offers an enhanced feature of samples in inference.

To improve the efficacy-performance exchange of preceding versions, they presented 3 features. (i) For the lightweight classification head, because of the better influence of the regression task, the overhead of the classifier head was diminished by presenting dual depth-wise separable convolutions. (ii) Spatial-channel decoupled down-sampling includes splitting the down-sampling procedure into dual stages: Initially, performing a point-wise convolution to fine-tune the channel size, accompanied by a depth-wise convolution to attain spatial down-sampling. (iii) A rank-guided block design is used to deal with the problem of redundancy in the deeper phases of methods produced by reiterating a similar structural block in each phase. The authors propose a Compact Inverted Block (CIB) architecture that utilizes depth-wise convolutions for spatial and point-wise convolutions for channel mixing.

Finally, to yield a precision-driven design of the method, authors accept dual approaches: (i) Large-kernel convolution in CIB inside deeper phases to increase the receptive area, thus improving the models' abilities. This was additionally applied in more minor method scales because of the rise of I/O latency and overhead. The authors present Partial *self-attention* (PSA) to decrease the computational cost of self-attention. PSA has been used in only one phase to prevent the unwarranted overhead related to self-attention, thus improving the model's abilities and performance. One portion is connected with a block called N_{PSA} . Which contains a



nature portfolio

In addition, the backbone fusion of feature extraction models such as CapsNet and InceptionV3 is employed to capture diverse spatial and contextual information. This integration is highly effective due to the complementary merits of both models. CapsNet is capable of preserving spatial hierarchies and capturing complex relationships between features, which enhances the capability of the model to generalize to unseen data, particularly in cases where orientation or viewpoint discrepancies are present. InceptionV3, on the contrary, is a deep network with multiple filter sizes, allowing it to capture both fine and coarse details efficiently. This hybrid approach utilizes CapsNet's dynamic routing mechanism to capture precise spatial dependencies while utilizing InceptionV3's deep layers for robust, multi-level feature extraction. This dual model strategy confirms high-quality feature extraction from varied data, outperforming conventional CNNs that often face difficulty with intrinsic spatial relationships. Furthermore, the incorporation provides a good balance between accuracy and computational efficiency.

CapsNet method

The basic notion of the CapsNet is to convert the neurons of conventional neural networks from scalar to vector and to apply vector as the system's input and output³¹. By passing from low-level to high-level capsule layers, not only may the loss of features be decreased, allowing the system to remove more abstract and subtle features. Simultaneously, the capsule activation is measured as the possibility of the feature presence, which permits CapsNet to deal with classification issues efficiently. Figure 4 represents the structure of the CapsNet model.

The major architecture of the CapsNet contains convolution layers, the first layers of the capsule, and the digit capsule. The convolution layers are applied to remove features from the input. The first capsule layer removes higher-dimension information from lower-level characteristics utilizing convolution and related models. It modifies the scalar neurons in conventional convolution systems into the type of capsule vector of a stated length. The dynamical routing model is the basic method of passing information from the layers of the low-level capsule to the high-level capsule. The particular procedure is as demonstrated:

- (1) Like convolutional operations, all neurons are multiplied through the weight. Unlike the convolution operation, capsules work on input vectors. So, the predicted vector is characterized as:

$$\widehat{u}_{j|i} = W_j u_i \quad (2)$$

Whereas $\widehat{u}_{j|i}$ signifies the forecast vector, W_j characterizes the weighted matrix of the j th iteration of the training, and u_i represents the i th higher layer vector.

- (2) To make the weighing amount of the predictive vectors to obtain the vector of the output, it is characterized as:

$$s_j = \sum_i c_{ij} \cdot \widehat{u}_{j|i} \quad (3)$$

Whereas s_j signifies the output vector, and c_{ij} denotes the coupling coefficient. The value of c_{ij} is upgraded by the dynamical routing model in training iteration, and their updated approach is associated with the $\text{Softmax}()$ activation function. The formulation is as shown:

$$c_{ij} = \frac{\exp(b_{ij})}{\sum_k \exp(b_{ik})} \quad (4)$$

Here, b_{ij} is primarily initialized as *zero* and acts as the temporary variable for calculation in following iterations. The updated formulation is as shown:

$$b_{ij} = b_{ij} + v_j \cdot \widehat{u}_{j|i} \quad (5)$$

- (3) The squash function has been applied to carry out a nonlinear mapping on the output vector to gain the last vector of the output. It is formulated as:

$$v_j = \text{squash}(s_j) = \frac{\|s_j\|^2}{1 + \|s_j\|^2} \frac{s_j}{\|s_j\|} \quad (6)$$

Now, v_j indicates the j th capsule's output vector and $\|\cdot\|$ signifies the L2-norm.

InceptionV3 method

The Inception-V3 approach, or GoogLeNet, is an extensively applied deep CNN (DCNN) structure³². It is well-known for its novel inception module, which integrates numerous convolution filters of dissimilar dimensions

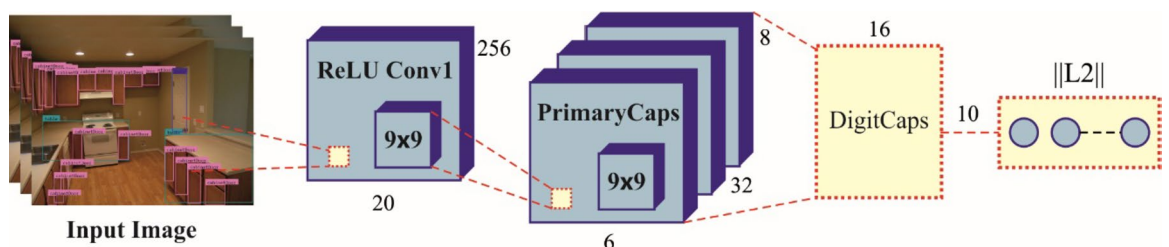


Fig. 4. Structure of CapsNet approach.

paralleled, permitting the system to take either global or local features effectively. This novel architecture aids in mitigating the problem of gradient vanishing and allows efficient feature extraction at different scales, which denotes significant benefits of Inception-V3. It uses numerous convolution kernels of dissimilar dimensions paralleled to handle input data, allowing the capturing of features through varied scales. This improves the capability of the network to observe structures and objects of dissimilar sizes. In addition, Inception-V3 combines a 1×1 convolution layer in the first module, which not only aids in preserving lower parameter amounts but also offers excellent representative abilities.

During this Inception-V3, the convolutional layer is its major element. This layer removes features from input data over convolution processes and utilizes weight and nonlinear conversions to this feature utilizing activation functions and convolutional filters. The mathematic model of the convolutional process is defined with the convolutional operator equation:

$$y_i = \sum_j x_i * w_{ij} + b_i \quad (7)$$

Meanwhile, x_i , b_j , and w_{ij} characterize the input, bias, and convolutional filter weights. Besides the convolutional process, the convolution layer utilizes an activation function to conduct nonlinear conversions on the output feature mapping. *ReLU* is a normally applied activation function that aids in mitigating gradient vanishing problems in training and enhances the learning ability of the networks. The function of *ReLU* is described as:

$$f(x) = \begin{cases} 0, & x < 0 \\ x, & x > 0 \end{cases} \quad (8)$$

The novel structure of Inception-V3 associates parallel convolutional filters, reducing the dimensions and effective resource management, which participates in its efficiency in processing different image datasets and dealing with challenges associated with computing efficiency. These features make Inception-V3 a general option in computer vision (CV) applications, mainly in object detection and image recognition tasks. The inception blocks A, B, and C are the first Inception-V3 modules. These modules are built using 1×1 , 3×3 , 7×7 , and 1×1 convolutional branches with max – pooling. The 1×1 convolutional branch is mainly applied for feature compression and dimensionality reduction, decreasing computational efficiency and overfitting prevention. The 3×3 convolutional branch takes attributes at a medium scale, helping identify features like edges, textures, and shapes. Conversely, the 7×7 convolutional branch is applied to capture large-scale characteristics, enabling the seizure of the complete architecture and contextual data of the features.

Classification process using the MHA-BiLSTM method

In the classification of the object detection process, the MHA-BiLSTM method is utilized³³. This method is particularly effective for classification tasks because it can capture temporal and contextual dependencies in sequential data. The component enables the model to process data forward and backward, capturing past and future context for better decision-making. The integration of MHA additionally improves this by allowing the model to concentrate on diverse crucial features concurrently, enhancing its capability to learn intrinsic patterns and associations. This combination enables the model to capture long-range dependencies and context-specific data, making it more robust than standard LSTM or attention-based models. Furthermore, MHA-BiLSTM can handle noisy or incomplete data better than conventional techniques, enhancing classification accuracy. Its flexibility and efficiency in handling sequential data make it superior to other classification approaches, mainly when dealing with complex and diverse datasets. Figure 5 indicates the MHA-BiLSTM methodology.

LSTM networks characterize a specific framework inside the RNN architecture. This network has been moved to mitigate the problems of explosion and gradient vanishing that frequently appear in conventional RNNs after processing expansive sequence data. This model allows LSTM models to take and maintain longer-term dependencies more efficiently. Bi-LSTM models increase the conventional LSTM architecture by incorporating bi-directional processing. Unlike normal LSTM models, which handle information in a unidirectional sequence from the previous to the future, Bi-LSTM models concurrently handle data in forward or backward directions. By combining the previous or following context, Bi-LSTM models can capture complex patterns and the dependencies of longer time inside sequential data.

Based on the Transformer method, the multiple-head attention mechanism presented characterizes significant progress over conventional attention mechanisms. It improves the attention method by concurrently using numerous attention *heads*, all with different parameters, toward input vectors. Considering the input vectors V (Value), K (Key), and Q (Query), every attention head works autonomously, and the calculation for the i th head is carried out as demonstrated.

Initially, linear transformations are used to get the V_i , Q_i , and K_i , for every attention head, signified as $Q_i = QW_Q^i$, $K_i = KW_K^i$, and $V_i = VW_V^i$, correspondingly. At the same time, W^i characterizes the learnable parameter matrices. The attention weights for i th head are computed as $A_j = \text{softmax} \left(\frac{Q_i K_i^T}{\sqrt{d_k}} \right) \cdot V_i$. At the same time, d_k denotes the dimensionalities of the K vectors, and separation by $\sqrt{d_k}$ is carried out to scale the values and stabilize the training procedure. At last, the outcomes from numerous heads are connected and pass over linear transformations to get the previous multi-head attention outcome, as provided by the succeeding equation:

$$MHA(Q, K, V) = \text{Concat}(A_1, A_2, \dots, A_n) W^O \quad (9)$$

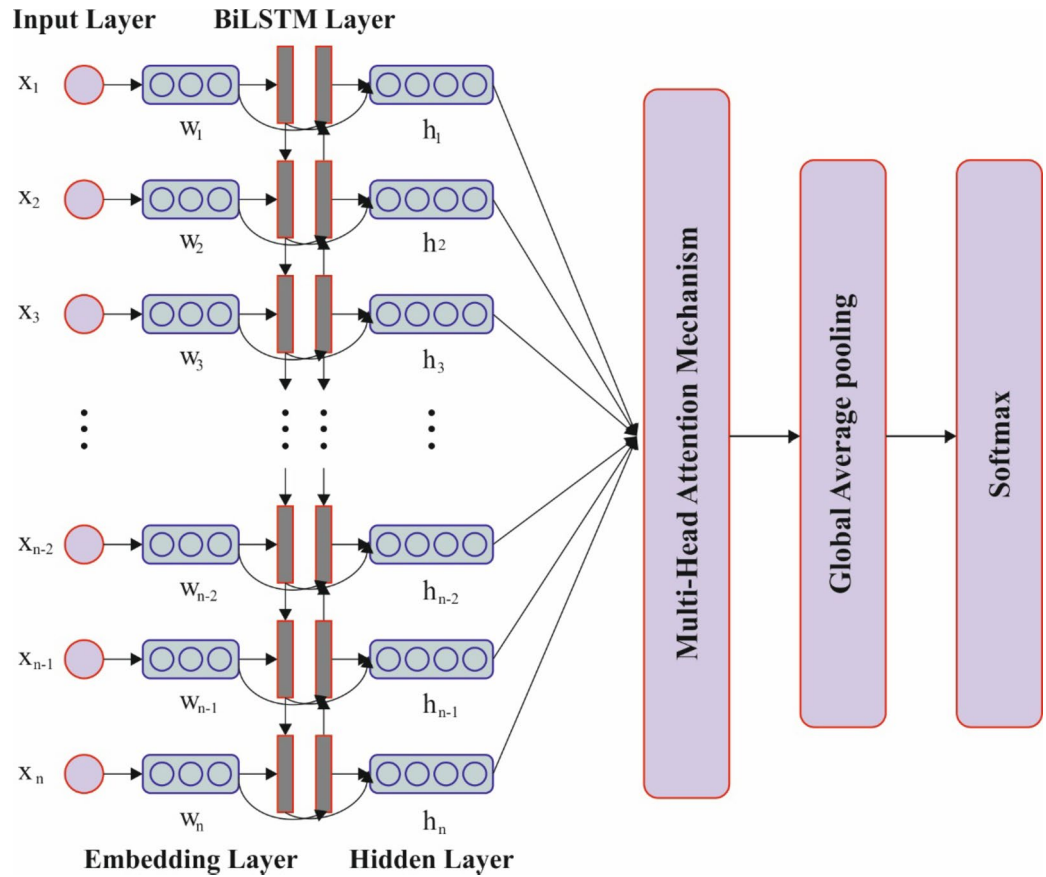


Fig. 5. Overall structure of the MHA-BiLSTM approach.

During Eq. (12), n characterizes the attention head counts, and W^O signifies the learnable parameter matrix. Moreover $W_i^Q \in \mathbb{R}^{d \times d_k}$, $W_i^K \in \mathbb{R}^{d \times d_k}$, $W_i^V \in \mathbb{R}^{d \times d_v}$, and $W^O \in \mathbb{R}^{n d_v \times d}$, whereas $d_k = d_v = d/n$.

The benefit of this structure is its ability to learn different attention patterns from different representative subdivisions, thus taking a richer and more different collection of feature relationships inside the input data. This considerably improves the model's performance and expressiveness. Nevertheless, Bi-LSTM methods face restrictions in computational complexity. When handling larger datasets owing to their incapability to carry out parallel calculations. The joined mould structure feeds the data dealt with by the Bi-LSTM hidden layers (HLs) into the multi-head attention system. These LSTM units in Bi-LSTM networks include 3 dissimilar gates: the input, the forget, and the output gates. The input gate outputs i_t , and the cell state of the candidate \tilde{C}_t is calculated utilizing the succeeding equations:

$$\begin{aligned} i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\ \tilde{C}_t &= \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \end{aligned} \quad (10)$$

The output f_t of the forget gate is calculated building on the present input x_t and the preceding HL h_{t-1} , as defined by the subsequent equation:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (11)$$

The forget gate's function is to select how many of the formerly deposited values must be residue from the state of the cell. The cell state C_t is upgraded as shown:

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \quad (12)$$

The output gate's output is calculated as shown:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (13)$$

The last HL h_t is computed as demonstrated:

$$h_t = o_t \odot \tanh(C_t) \quad (14)$$

During Eqs. (10)–(14), σ signifies the functions of the sigmoid, W_* characterizes the weighted matrices for the gates, b_* is mentioned as the biases related to the candidate and gates states, x_i specifies the present input, h_{t-1} means the HL from the preceding time step, and \odot represents element-to-element multiplications.

The calculation equations for the forward and reverse layers of the LSTM in a Bi-LSTM technique are equivalent. The input and output computations for the Bi-LSTM technique are as shown:

$$\begin{aligned} f_t^* &= \sigma(W_f^* \cdot [h_{t-1}^*, x_t] + b_f^*) \\ i_t^* &= \sigma(W_i^* \cdot [h_{t-1}^*, x_t] + b_i^*) \end{aligned} \quad (15)$$

Equation (15), $*$ characterizes element-to-element multiplication, whereas \odot represents the element-to-element multiplication for the forward and backward LSTM. In particular, a and b are associated with the “−” and “+” sign. The final output of the Bi-LSTM technique is completed by connecting the HL of forward and backwards, indicated as $h_t = [h_t^a, h_t^b]$. This connection incorporates the temporal features obtained in either direction, improving the model’s capability to take composite dependence inside the information.

Hyperparameter tuning using the PIO model

Finally, the hyperparameter tuning process is performed through the PIO approach to advance the classification performance of the MHA-BiLSTM model³⁴. This model is effectual due to its unique inspiration from the foraging behaviour of pigeons, which enables efficient global search and exploration of the solution space. PIO outperforms at balancing exploration and exploitation, making it highly appropriate for finding optimal hyperparameters in complex, high-dimensional models. Unlike conventional gradient-based methods, which may get stuck in local optima, PIO utilizes a population-based approach that improves the robustness of the search process. Furthermore, PIO is computationally efficient and can handle non-differentiable, noisy, or multi-modal objective functions, which is standard in real-world ML tasks. Its adaptability and capability to escape local minima make it superior to other optimization techniques, such as grid or random search, resulting in enhanced model performance and accuracy. Figure 6 describes the steps involved in the PIO method.

Pigeons returning home can discover their way back home, and it takes a very short time to move. The PIO model is according to (i) a map and compass operator and (ii) a landmark operator. The map and compass operator have been employed in the initial stages of returning home for pigeons. The map and compass will continue to direct pigeons to fly until they reach their home place’s neighbourhood. At present, the position of the map and compass operator is reduced, and the landmarks play an essential part in pigeon flight. The

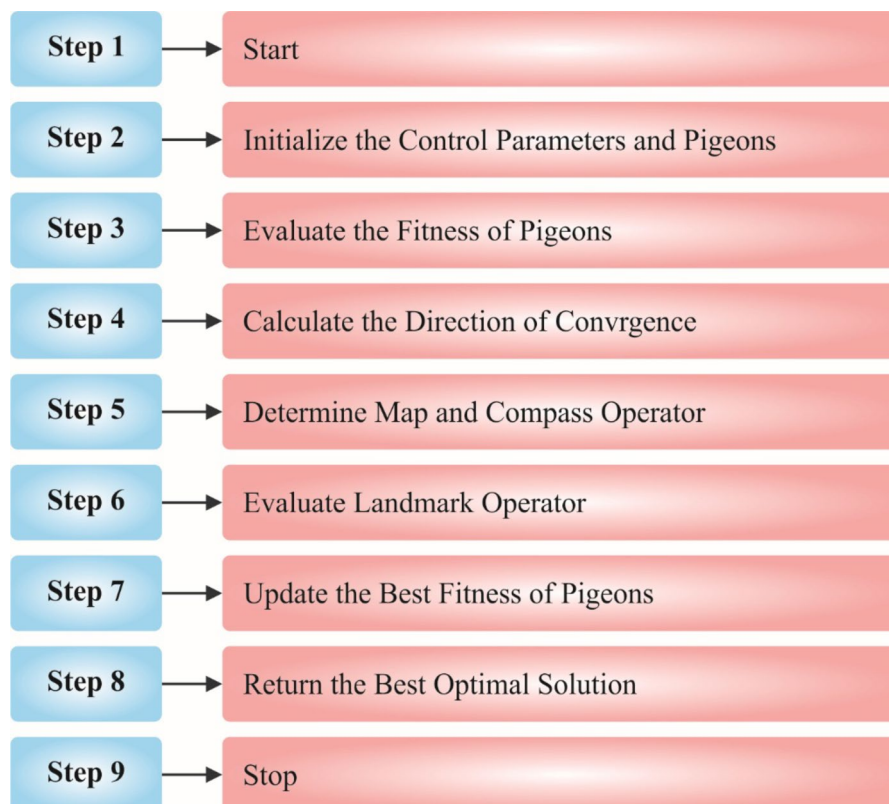


Fig. 6. Steps involved in the PIO model.

optimizer process of the model depends on the pigeons' returning home behaviour. The location of all pigeons is the candidate's solution to the considered mathematic issue. In continuous flight, respective locations and the movement direction of the pigeons are continuously fine-tuned to gain the optimum solution.

Map and Compass Operator: Past work has exposed pigeons' remark of the magnetic fields and made a map in their mind. Additionally, the map and compass operator have been applied to fine-tune the location of all pigeons based on the magnetic field map. All pigeons have a different location and speed. The pigeon's position, Pos , characterizes a candidate solution, and the speed, Vel , represents the movement tendency of the pigeon at the following iteration. The Pos and Vel magnitudes are established by the size of the problem to be enhanced. The mathematical representation of the map and compass operator is stated in Eqs. (16) and (17).

$$Vel_i^t = Vel_i e^{-R \times t} + rand() \cdot (Pos_g^{t-1} - Pos_i^{t-1}) \quad (16)$$

R refers to the factor associated with the map and compass operator, modified according to various problems. The parameter $rand$ is comprised of randomly generated values among $(0,1)$. Pos_g^{t-1} denotes the finest location, the global optimum location at iteration $t-1$.

$$Pos_i^t = Pos_i^{t-1} + Vel_i^t \quad (17)$$

t refers to iteration count. Pos and Vel represent the velocity and position of pigeon i in the present iteration.

Landmark Operator: Some studies have exposed that pigeons will get landmark information after flying. A landmark might be a river, a building, or a tree. During the landmark operator in PIO, the inner of the pigeon circle at the middle point signifies all generations of brilliant solutions, and the outer of the circle signifies the pigeons that can't identify the optimum travelling route. These pigeons should be unnoticed. The pigeons in the circle will endure for iteration after the pigeon in the middle of the circle.

The pigeons outside the circle cannot identify the optimum travelling route and are unnoticed in an iterative procedure. Hence, half of the pigeons are rejected in every iteration. The pigeons in the circle can encourage the convergence of the model and are comprised in the following iterations.

$$N_p^t = ceil\left(\frac{N_p^{t-1}}{2}\right) \quad (18)$$

Here, N_p denotes pigeon counts in the population. At t th iteration, the changes in the pigeon count in the population are gained by Eq. (18). Simultaneously, $ceil(x)$ characterizes a rounding process that converts the value of x into the integer neighbouring to x . Assume that all pigeons in the population can fly straightly towards home. Formerly, the location of pigeon i is upgraded, as exposed in Eq. (19).

$$Pos_i^t = Pos_i^{t-1} + rand() \cdot (Pos_c^t - Pos_i^{t-1}) \quad (19)$$

Here, Pos_c^t symbolizes the virtual or real centre pigeon location at t iteration, and Eq. (20) is described as:

$$Pos_c^t = \frac{\sum_i^N (Pos_i(t) \cdot Fitness(Pos_i(t)))}{N_p(t) \sum_i^{N_p} Fitness(Pos_i(t))} \quad (20)$$

$Fitness(x)$ is derived from the fitness assessment standard for all pigeon locations, and this function varies from maximal to minimal. Equation (21) describes $Fitness(x)$ as exposed.

$$Fitness(Pos_i^t) \begin{cases} \frac{1}{fitness(Pos_i^t + \epsilon)} & \text{Minimum problem} \\ fitness(Pos_i^t) & \text{Maximum problem} \end{cases} \quad (21)$$

The PIO technique raises a fitness function (FF) to enhance the classifier's performance. It determines a positive numeral to indicate the improved outcome of the candidate solution. Here, the classification error rate reduction has been measured as FF. Its mathematical formulation is expressed in Eq. (22).

$$fitness(x_i) = ClassifierErrorRate(x_i) = \frac{No. of misclassified instances}{Total no. of instances} \times 100 \quad (22)$$

Experimental analysis

The performance evaluation of the HAODVIP-ADL approach is studied under the Indoor objects' detection dataset³⁵. The dataset contains 6642 counts under 10 objects, as portrayed in Table 1. Figure 7 demonstrates the sample images, while object detection images are depicted in Fig. 8.

Figure 9 establishes the confusion matrices created by the HAODVIP-ADL technique under different epochs. The outcomes recognize that the HAODVIP-ADL methodology has effectual detection and recognition of all ten class labels specifically.

The object detection of the HAODVIP-ADL approach is determined under distinct epochs in Table 2; Fig. 10. The table values state that the HAODVIP-ADL approach correctly recognized all the samples. On 500 epochs, the HAODVIP-ADL methodology provides an average $accu_y$ of 99.26%, $prec_n$ of 94.06%, $reca_l$ of 76.67%,

| Objects | Count |
|-------------------|-------|
| Door | 562 |
| Cabinet door | 3890 |
| Refrigerator door | 879 |
| Window | 482 |
| Chair | 223 |
| Table | 248 |
| Cabinet | 208 |
| Couch | 24 |
| Opened door | 90 |
| Pole | 36 |
| Total | 6642 |

Table 1. Details of the dataset.

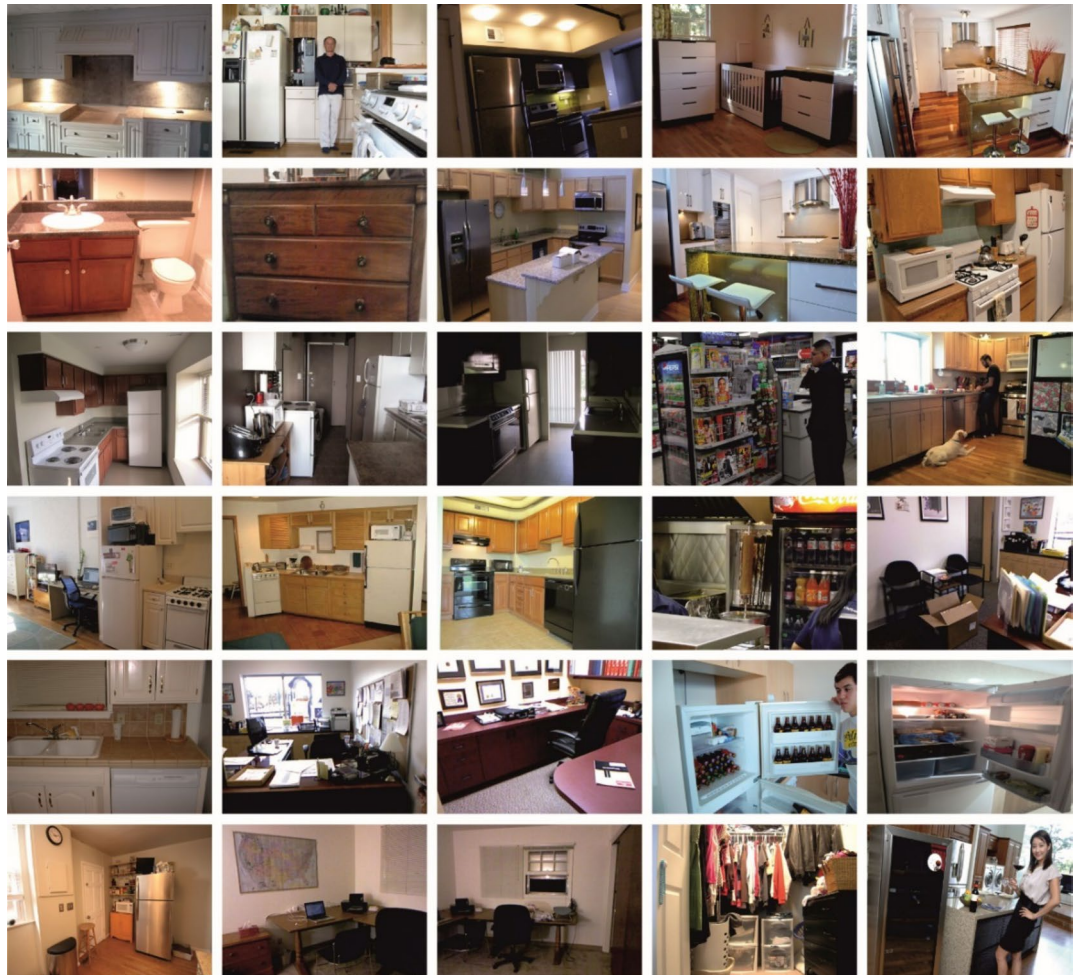


Fig. 7. Sample images.

F_{score} of 80.57%, and AUC_{score} of 88.06%. Besides, on 1000 epochs, the HAODVIP-ADL methodology offers an average $accu_y$ of 99.32%, $prec_n$ of 93.86%, $reca_l$ of 78.99%, F_{score} of 83.03%, and AUC_{score} of 89.26%. Moreover, on 1500 epochs, the HAODVIP-ADL approach delivers an average $accu_y$ of 99.39%, $prec_n$ of 94.89%, $reca_l$ of 79.79%, F_{score} of 83.83%, and AUC_{score} of 89.69%. Also, on 2500 epochs, the HAODVIP-ADL approach provides an average $accu_y$ of 99.54%, $prec_n$ of 95.28%, $reca_l$ of 84.06%, F_{score} of 87.68%, and AUC_{score} of 91.87%. At last, on 3000 epochs, the HAODVIP-ADL approach offers an average $accu_y$ of 99.74%, $prec_n$ of 96.41%, $reca_l$ of 91.41%, F_{score} of 93.42%, and AUC_{score} of 95.62%.



Fig. 8. Object detection images.

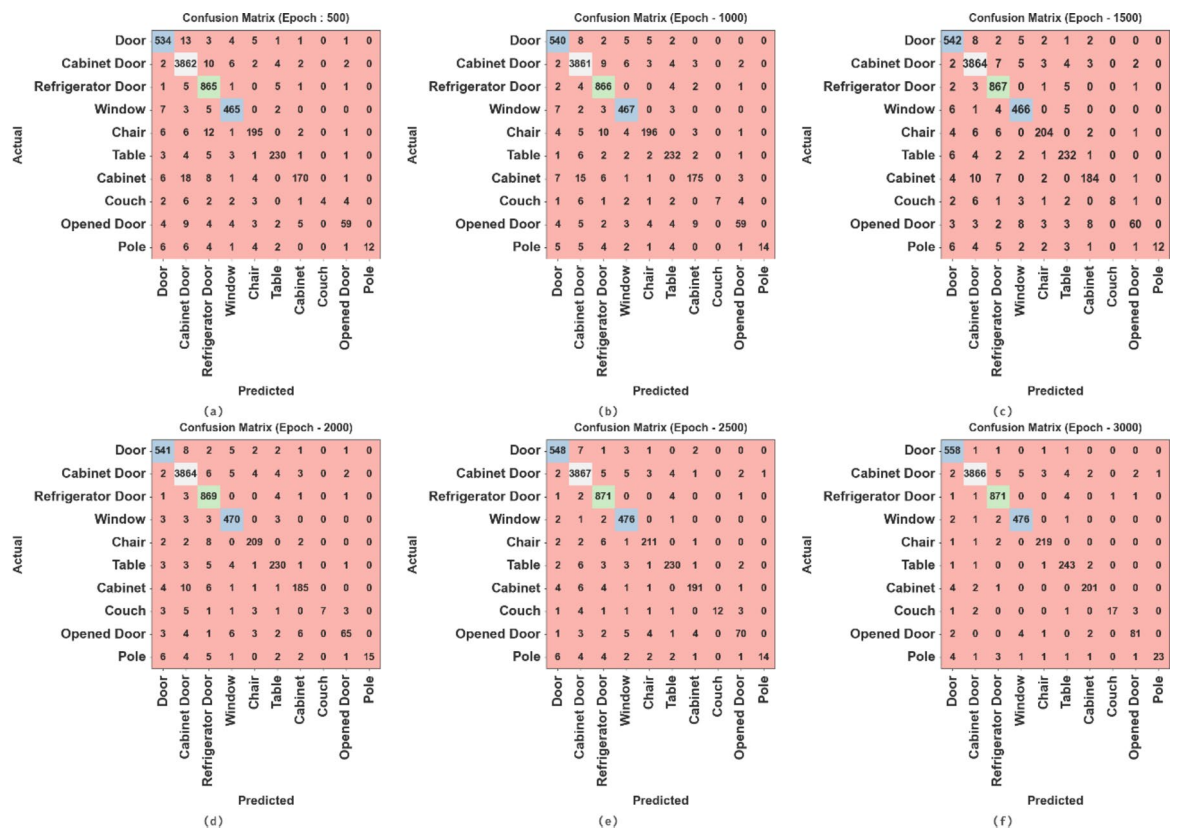


Fig. 9. Confusion matrix of HAODVIP-ADL methodology (a–f) Epochs 500–3000.

| Class labels | $Accu_y$ | $Prec_n$ | $Recall_l$ | F_{score} | AUC_{score} |
|-------------------|----------|----------|------------|-------------|---------------|
| Epoch – 500 | | | | | |
| Door | 99.02 | 93.52 | 95.02 | 94.26 | 97.20 |
| Cabinet door | 98.52 | 98.22 | 99.28 | 98.75 | 98.37 |
| Refrigerator door | 98.99 | 94.23 | 98.41 | 96.27 | 98.74 |
| Window | 99.40 | 95.29 | 96.47 | 95.88 | 98.05 |
| Chair | 99.25 | 89.86 | 87.44 | 88.64 | 93.55 |
| Table | 99.49 | 93.50 | 92.74 | 93.12 | 96.25 |
| Cabinet | 99.23 | 92.90 | 81.73 | 86.96 | 90.76 |
| Couch | 99.70 | 100.00 | 16.67 | 28.57 | 58.33 |
| Opened door | 99.35 | 83.10 | 65.56 | 73.29 | 82.69 |
| Pole | 99.64 | 100.00 | 33.33 | 50.00 | 66.67 |
| Average | 99.26 | 94.06 | 76.67 | 80.57 | 88.06 |
| Epoch – 1000 | | | | | |
| Door | 99.17 | 94.24 | 96.09 | 95.15 | 97.77 |
| Cabinet door | 98.72 | 98.57 | 99.25 | 98.91 | 98.61 |
| Refrigerator door | 99.22 | 95.69 | 98.52 | 97.09 | 98.92 |
| Window | 99.40 | 94.92 | 96.89 | 95.89 | 98.24 |
| Chair | 99.34 | 92.02 | 87.89 | 89.91 | 93.81 |
| Table | 99.41 | 90.98 | 93.55 | 92.25 | 96.59 |
| Cabinet | 99.22 | 90.21 | 84.13 | 87.06 | 91.92 |
| Couch | 99.74 | 100.00 | 29.17 | 45.16 | 64.58 |
| Opened door | 99.34 | 81.94 | 65.56 | 72.84 | 82.68 |
| Pole | 99.67 | 100.00 | 38.89 | 56.00 | 69.44 |
| Average | 99.32 | 93.86 | 78.99 | 83.03 | 89.26 |
| Epoch – 1500 | | | | | |
| Door | 99.17 | 93.93 | 96.44 | 95.17 | 97.93 |
| Cabinet door | 98.93 | 98.85 | 99.33 | 99.09 | 98.85 |
| Refrigerator door | 99.28 | 96.01 | 98.63 | 97.31 | 99.01 |
| Window | 99.38 | 94.91 | 96.68 | 95.79 | 98.14 |
| Chair | 99.49 | 93.15 | 91.48 | 92.31 | 95.62 |
| Table | 99.41 | 90.98 | 93.55 | 92.25 | 96.59 |
| Cabinet | 99.38 | 91.54 | 88.46 | 89.98 | 94.10 |
| Couch | 99.76 | 100.00 | 33.33 | 50.00 | 66.67 |
| Opened door | 99.44 | 89.55 | 66.67 | 76.43 | 83.28 |
| Pole | 99.64 | 100.00 | 33.33 | 50.00 | 66.67 |
| Average | 99.39 | 94.89 | 79.79 | 83.83 | 89.69 |
| Epoch – 2000 | | | | | |
| Door | 99.28 | 95.25 | 96.26 | 95.75 | 97.91 |
| Cabinet door | 98.98 | 98.92 | 99.33 | 99.13 | 98.90 |
| Refrigerator door | 99.29 | 95.92 | 98.86 | 97.37 | 99.11 |
| Window | 99.47 | 95.33 | 97.51 | 96.41 | 98.57 |
| Chair | 99.58 | 93.72 | 93.72 | 93.72 | 96.75 |
| Table | 99.44 | 92.37 | 92.74 | 92.56 | 96.22 |
| Cabinet | 99.41 | 92.04 | 88.94 | 90.46 | 94.35 |
| Couch | 99.74 | 100.00 | 29.17 | 45.16 | 64.58 |
| Opened door | 99.49 | 87.84 | 72.22 | 79.27 | 86.04 |
| Pole | 99.68 | 100.00 | 41.67 | 58.82 | 70.83 |
| Average | 99.44 | 95.14 | 81.04 | 84.87 | 90.33 |
| Epoch – 2500 | | | | | |
| Door | 99.47 | 96.31 | 97.51 | 96.91 | 98.58 |
| Cabinet door | 99.13 | 99.10 | 99.41 | 99.26 | 99.07 |
| Refrigerator door | 99.46 | 96.89 | 99.09 | 97.98 | 99.30 |
| Window | 99.59 | 95.77 | 98.76 | 97.24 | 99.21 |
| Chair | 99.62 | 94.20 | 94.62 | 94.41 | 97.21 |
| Table | 99.53 | 94.65 | 92.74 | 93.69 | 96.27 |
| Cabinet | 99.59 | 95.02 | 91.83 | 93.40 | 95.84 |
| Continued | | | | | |

| Class labels | $Accu_y$ | $Prec_n$ | $Reca_l$ | F_{score} | AUC_{score} |
|-------------------|----------|----------|----------|-------------|---------------|
| Couch | 99.82 | 100.00 | 50.00 | 66.67 | 75.00 |
| Opened door | 99.55 | 87.50 | 77.78 | 82.35 | 88.81 |
| Pole | 99.65 | 93.33 | 38.89 | 54.90 | 69.44 |
| Average | 99.54 | 95.28 | 84.06 | 87.68 | 91.87 |
| Epoch - 3000 | | | | | |
| Door | 99.67 | 96.88 | 99.29 | 98.07 | 99.50 |
| Cabinet door | 99.49 | 99.74 | 99.38 | 99.56 | 99.51 |
| Refrigerator door | 99.67 | 98.42 | 99.09 | 98.75 | 99.42 |
| Window | 99.76 | 97.94 | 98.76 | 98.35 | 99.30 |
| Chair | 99.83 | 96.90 | 98.21 | 97.55 | 99.05 |
| Table | 99.74 | 95.29 | 97.98 | 96.62 | 98.90 |
| Cabinet | 99.79 | 96.63 | 96.63 | 96.63 | 98.26 |
| Couch | 99.88 | 94.44 | 70.83 | 80.95 | 85.41 |
| Opened door | 99.76 | 92.05 | 90.00 | 91.01 | 94.95 |
| Pole | 99.79 | 95.83 | 63.89 | 76.67 | 81.94 |
| Average | 99.74 | 96.41 | 91.41 | 93.42 | 95.62 |

Table 2. Object detection of HAODVIP-ADL technique under dissimilar epochs.

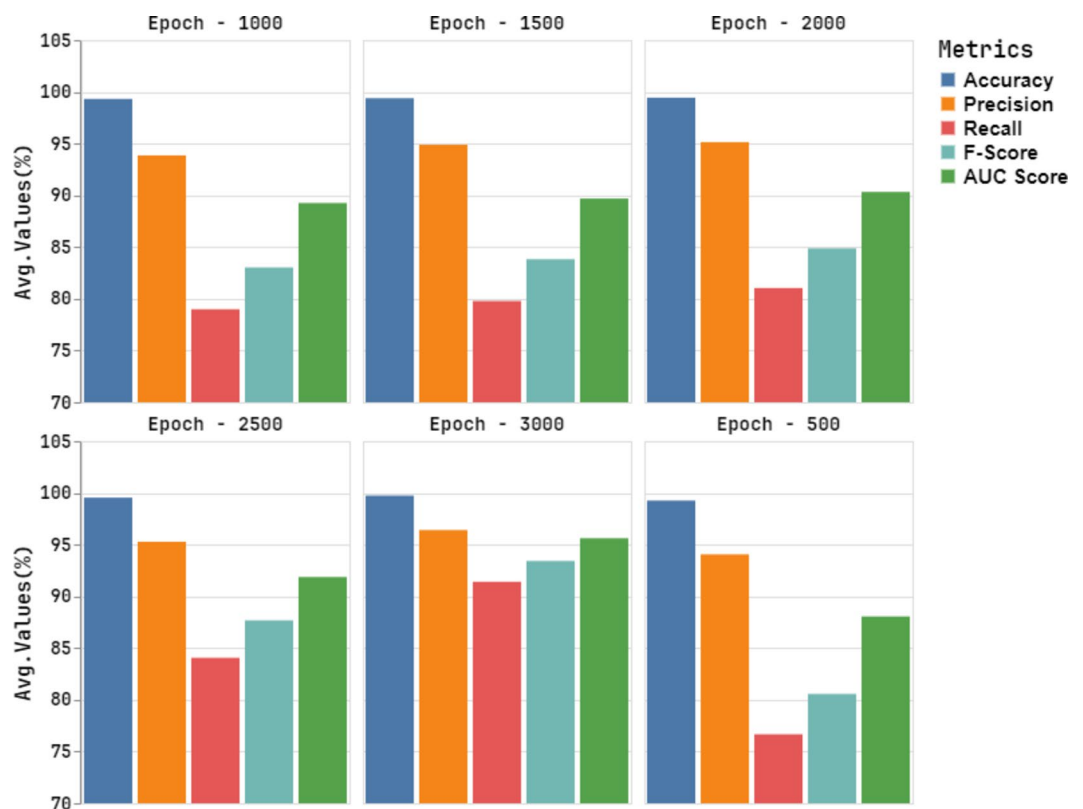


Fig. 10. Average of HAODVIP-ADL technique under distinct epochs.

Figure 11 illustrates the training (TRA) $accu_y$ and validation (VAL) $accu_y$ outcomes of the HAODVIP-ADL methodology under different epochs. The $accu_y$ analysis is calculated over the range of 0–3000 epochs. The figure highlights that the TRA and VAL $accu_y$ analysis demonstrates an increasing trend which informed the capacity of the HAODVIP-ADL methodology with superior performance across various iterations. At the same time, the TRA and VAL $accu_y$ get closer across the epochs, which specifies inferior overfitting and exhibitions maximal outcomes of the HAODVIP-ADL approach, guaranteeing constant prediction on hidden samples.

Figure 12 illustrates the TRA loss (TRALOS) and VAL loss (VALLOS) curves of the HAODVIP-ADL method under distinct epochs. The loss values are calculated across an interval of 0–3000 epochs. The TRALOS and VALLOS values exemplify a reducing trend, informing the capacity of the HAODVIP-ADL method to balance a

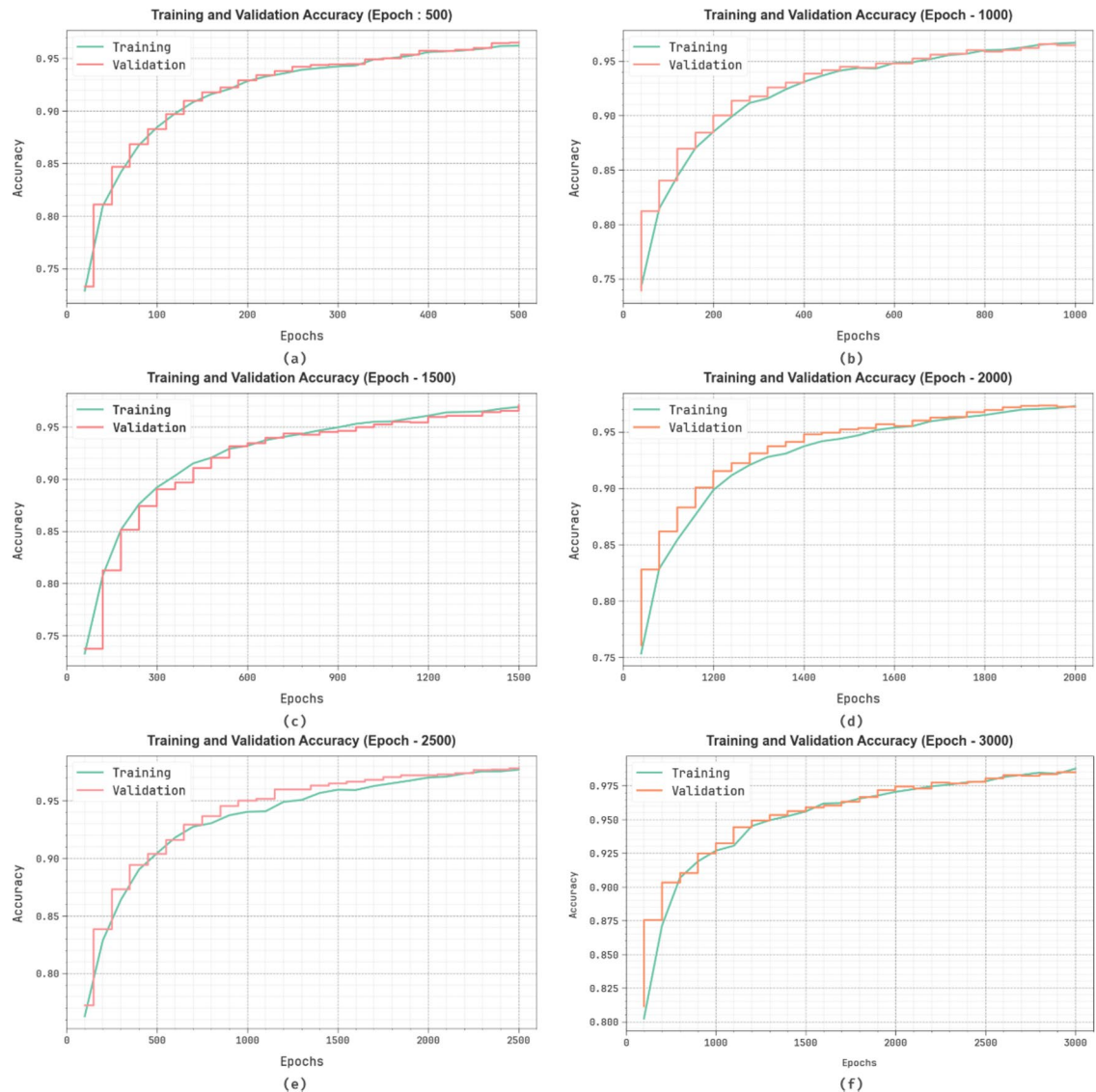


Fig. 11. $Accu_y$ curve of HAODVIP-ADL approach (a–f) Epochs 500–3000.

trade-off between data fitting and simplification. The continuous reduction in loss values assures the maximum outcomes of the HAODVIP-ADL technique and tunes the prediction results over time.

Table 3; Fig. 13 compare the HAODVIP-ADL approach with the existing methodologies^{20,36–38}. The results highlight that the IOD155 + tfidf, IOD90 + tfidf, Xception + tfidf, YOLOv5n, YOLOv5m, Yolo Tiny, Yolo V3, ResNET101, Support Vector Machine (SVM), and Linear Discriminant Analysis (LDA) methodologies have reported worse performance. Meanwhile, the Retina Net technique has achieved closer outcomes. Furthermore, the HAODVIP-ADL technique reported higher performance with superior $prec_n$, $recal$, $accu_y$, and $F_{measure}$ of 96.41%, 91.41%, 99.74%, and 93.42%, correspondingly.

In Table 4; Fig. 14, the comparative results of the HAODVIP-ADL approach are identified in mAP. The results imply that the HAODVIP-ADL model gets better performance. Based on mAP, the HAODVIP-ADL model offers a maximal mAP of 94.76%. In contrast, the IOD155 + tfidf, IOD90 + tfidf, Xception + tfidf, YOLOv5n, YOLOv5m, Retina Net, Yolo Tiny and Yolo V3, ResNET101, SVM, and LDA approaches reach minimal mAP of 86.70%, 88.20%, 92.10%, 86.57%, 88.81%, 91.39%, 88.37%, 91.33%, 92.00%, 90.11%, and 91.32%, respectively.

The mean IoU result of the HAODVIP-ADL model is denoted in Table 5; Fig. 15. The outcomes suggest that the HAODVIP-ADL methodology achieves effective results. The existing methods IOD155 + tfidf, IOD90 + tfidf, Xception + tfidf, YOLOv5n, YOLOv5m, Retina Net, Yolo Tiny, Yolo V3, ResNET101, SVM, and LDA methods accomplish lower mean IoU values of 67.33%, 66.64%, 65.62%, 67.16%, 64.28%, 43.31%, 69.19%, 68.22%, 69.00%, 68.49%, and 70.45%, correspondingly. Simultaneously, the proposed HAODVIP-ADL model gains a maximum mean IoU value of 76.06%.

Table 6; Fig. 16 illustrate the computational time (CT) analysis of the HAODVIP-ADL method with existing techniques. The methods span from image processing models such as YOLO variants such as YOLOv5n, YOLOv5m, YOLO Tiny, YOLOv3, and RetinaNet to conventional ML methods like SVM and LDA, and

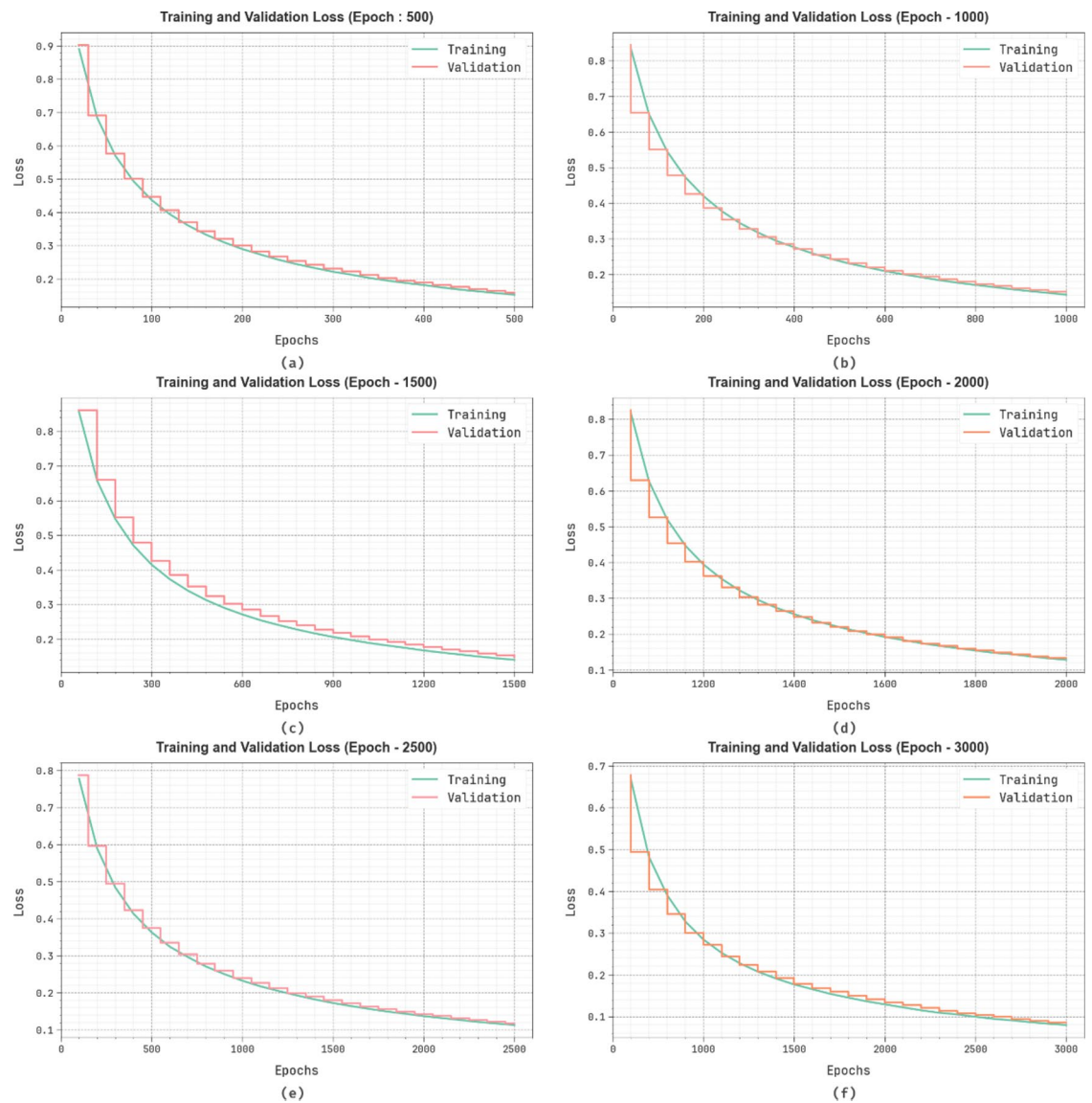


Fig. 12. Loss analysis of HAODVIP-ADL approach (a–f) Epochs 500–3000.

| Techniques | $Accu_y$ | $Prec_n$ | $Recal$ | $F_{measure}$ |
|------------------|----------|----------|---------|---------------|
| IOD155 + tfidf | 83.38 | 87.52 | 88.12 | 88.28 |
| IOD90 + tfidf | 83.63 | 87.61 | 87.51 | 88.70 |
| Xception + tfidf | 71.41 | 88.97 | 89.09 | 86.73 |
| YOLOv5n | 93.00 | 87.94 | 90.05 | 88.91 |
| YOLOv5m | 93.60 | 94.03 | 87.82 | 89.84 |
| Retina Net | 98.60 | 95.27 | 91.55 | 92.55 |
| Yolo Tiny | 80.70 | 88.10 | 88.64 | 87.85 |
| Yolo V3 | 77.60 | 88.92 | 87.82 | 87.10 |
| ResNET101 | 86.13 | 90.12 | 91.30 | 91.69 |
| SVM | 86.83 | 91.26 | 90.36 | 91.75 |
| LDA | 94.37 | 93.15 | 91.02 | 90.59 |
| HAODVIP-ADL | 99.74 | 96.41 | 91.41 | 93.42 |

Table 3. Comparative outcomes of the HAODVIP-ADL approach with existing techniques^{20,36–38}.

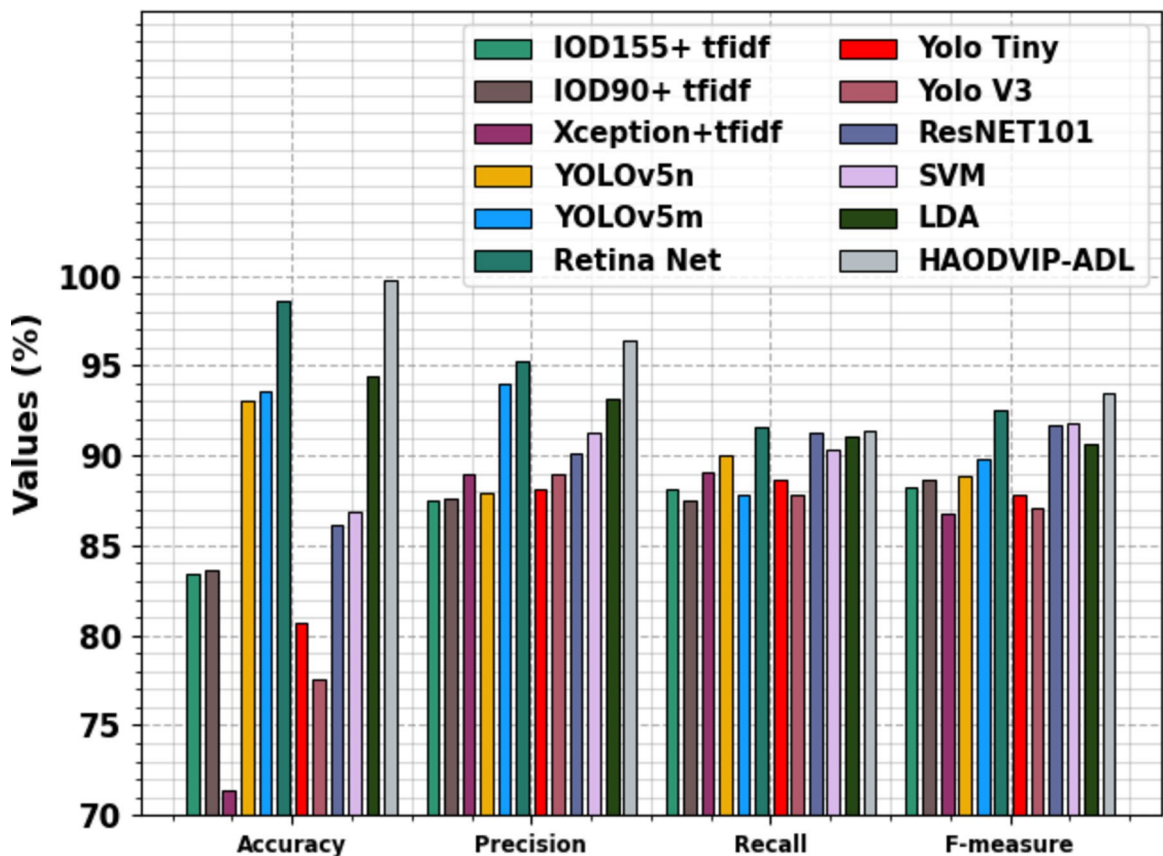


Fig. 13. Comparative analysis of HAODVIP-ADL approach with existing techniques.

| Techniques | mAP@0.5 (%) |
|------------------|-------------|
| IOD155 + tfidf | 86.70 |
| IOD90 + tfidf | 88.20 |
| Xception + tfidf | 92.10 |
| YOLOv5n | 86.57 |
| YOLOv5m | 88.81 |
| Retina Net | 91.39 |
| Yolo Tiny | 88.37 |
| Yolo V3 | 91.33 |
| ResNET101 | 92.00 |
| SVM | 90.11 |
| LDA | 91.32 |
| HAODVIP-ADL | 94.76 |

Table 4. mAP result of HAODVIP-ADL technique with recent models.

DL models such as Xception and ResNet101. The time ranges from the quickest, HAODVIP-ADL, at 9.00 s to the slowest, LDA, at 23.22 s. Notably, YOLOv5m and ResNet101 exhibit relatively fast CTs of 10.56 s and 14.25 s, respectively, while methods such as IOD155 plus tfidf at 13.91 s and Xception plus tfidf at 17.39 s show good trade-offs between speed and feature extraction. Overall, HAODVIP-ADL outperforms other methods, suggesting an effectual approach for computational tasks.

Conclusion

In this study, a HAODVIP-ADL model is presented. The major intention of the HAODVIP-ADL model is to deliver a reliable and precise object detection system that supports a visually impaired person in navigating their surroundings safely and effectively. The presented HAODVIP-ADL method initially employed BF for the image pre-processing stage to reduce noise while preserving edges for clarity. For object detection, the HAODVIP-ADL

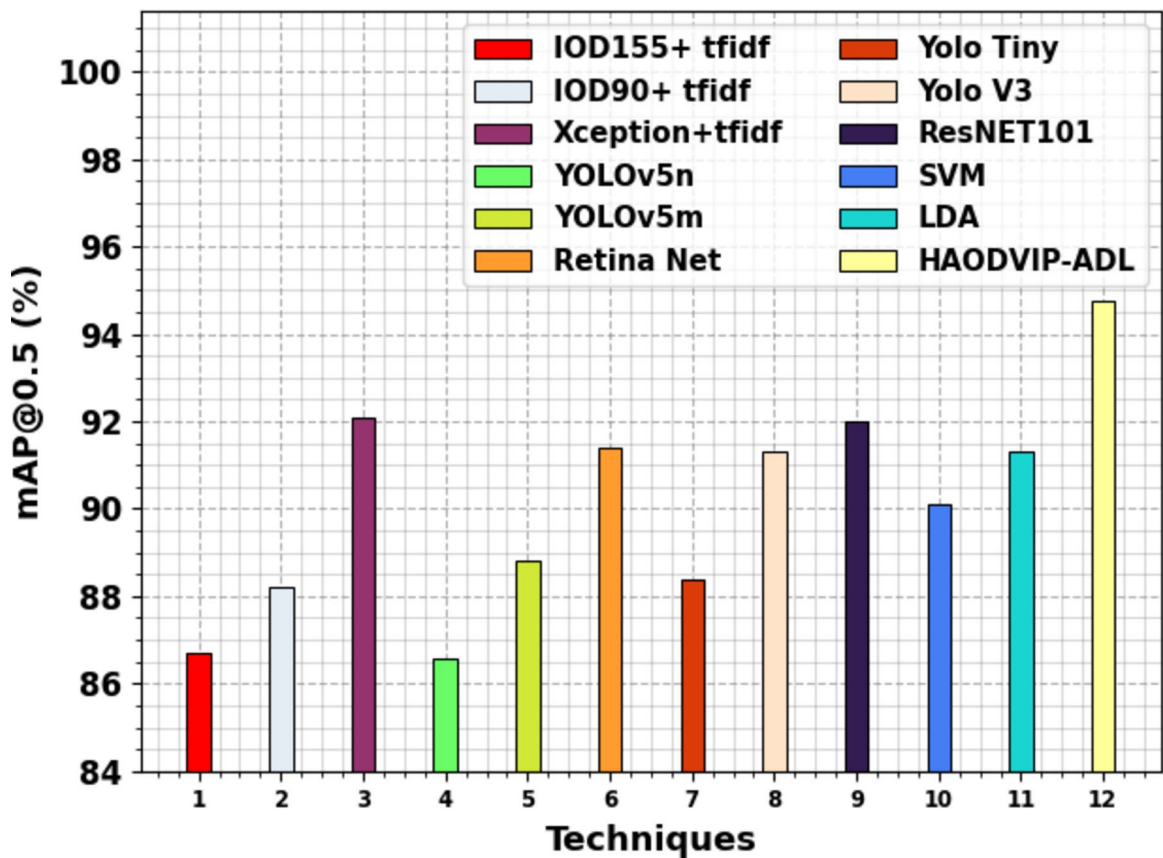


Fig. 14. mAP outcome of HAODVIP-ADL technique with recent models.

| Techniques | Mean IoU (%) |
|------------------|--------------|
| IOD155 + tfidf | 67.33 |
| IOD90 + tfidf | 66.64 |
| Xception + tfidf | 65.62 |
| YOLOv5n | 67.16 |
| YOLOv5m | 64.28 |
| Retina Net | 43.31 |
| Yolo Tiny | 69.19 |
| Yolo V3 | 68.22 |
| ResNET101 | 69.00 |
| SVM | 68.49 |
| LDA | 70.45 |
| HAODVIP-ADL | 76.06 |

Table 5. Mean IoU outcome of HAODVIP-ADL technique with recent models.

model utilized the YOLOV10 framework. In addition, the backbone fusion of feature extraction models such as CapsNet and InceptionV3 is employed to capture diverse spatial and contextual information. The MHA-BiLSTM model is utilized to classify the object detection process. Finally, the hyperparameter tuning process is performed through the PIO model to advance the classification performance of the MHA-BiLSTM method. The experimental results of the HAODVIP-ADL method are analyzed, and the outcomes are evaluated using the Indoor Objects Detection dataset. The experimental validation of the HAODVIP-ADL method portrayed a superior accuracy value of 99.74% over the existing methods. The limitations of the HAODVIP-ADL method comprise the dependence on a limited set of data sources, which may not fully represent the diversity and complexity of real-world scenarios. Moreover, computational efficiency could be additionally optimized for real-time applications, as the model may still need crucial processing power in resource-constrained environments. The study also lacks comprehensive testing across a wide range of datasets, which could affect the generalization of the proposed approach. Furthermore, the existing methodology does not fully address potential challenges

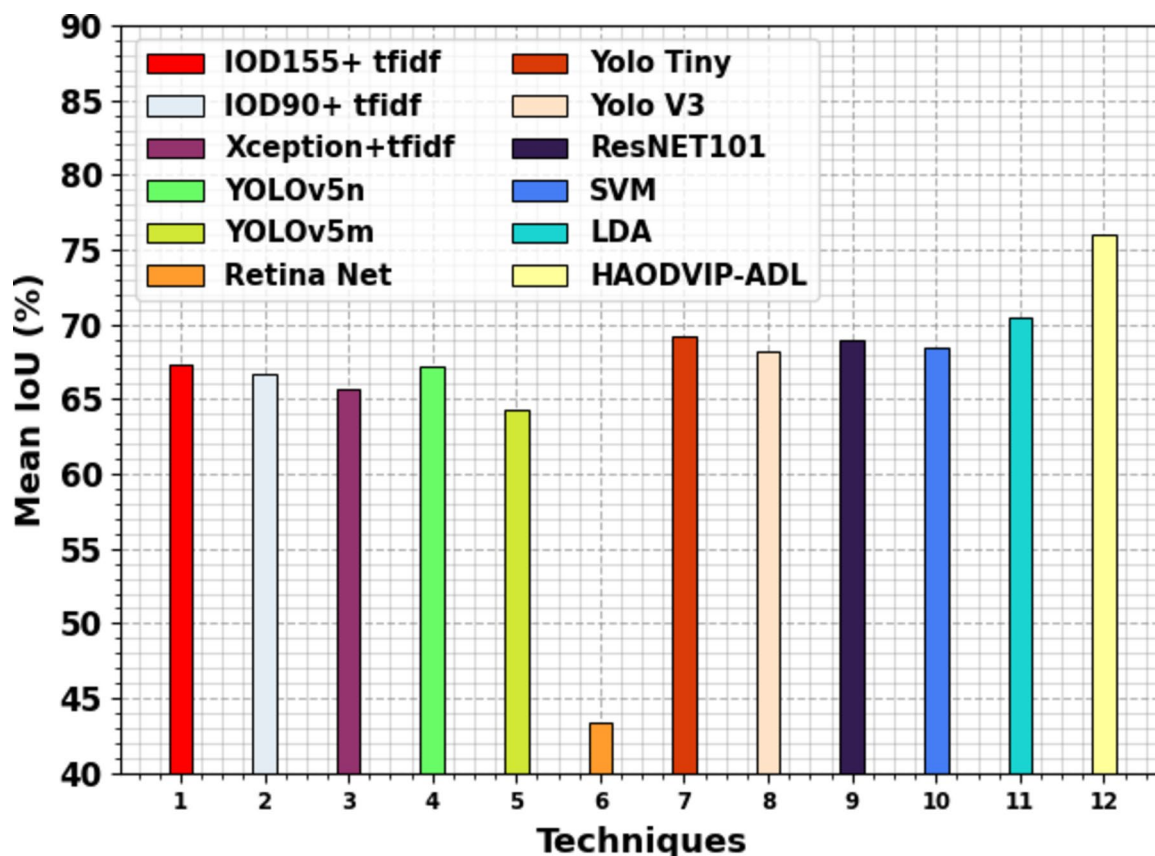


Fig. 15. Mean IoU outcome of HAODVIP-ADL technique with recent models.

| Techniques | CT (s) |
|-----------------|--------|
| IOD155+ tfidf | 13.91 |
| IOD90+ tfidf | 22.26 |
| Xception+ tfidf | 17.39 |
| YOLOv5n | 17.97 |
| YOLOv5m | 10.56 |
| Retina Net | 21.33 |
| Yolo Tiny | 19.78 |
| Yolo V3 | 22.46 |
| ResNET101 | 14.25 |
| SVM | 14.22 |
| LDA | 23.22 |
| HAODVIP-ADL | 9.00 |

Table 6. CT analysis of HAODVIP-ADL approach with existing techniques.

related to model interpretability and explainability. Future work can concentrate on improving the robustness of the model, improving its scalability for massive datasets, and integrating techniques for model transparency to provide more precise insights into decision-making processes. Expanding the framework to comprise domain-specific datasets and incorporating multimodal data could also improve performance. Finally, real-time deployment and practical testing would be crucial for evaluating its applicability.

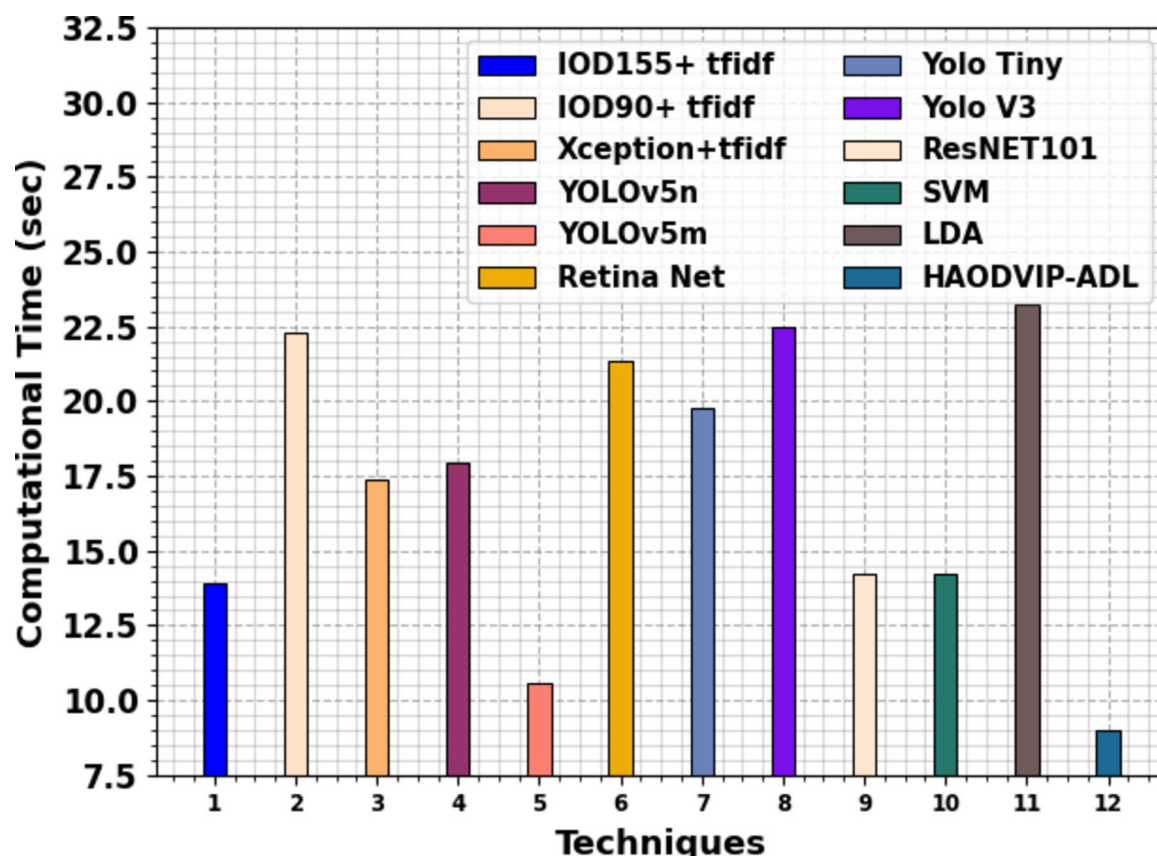


Fig. 16. CT analysis of HAODVIP-ADL approach with existing techniques.

Data availability

The data supporting this study's findings are openly available in the Kaggle repository at <https://www.kaggle.com/datasets/thebordin/indoor-object-detection>, reference number³⁵.

Received: 14 December 2024; Accepted: 26 February 2025

Published online: 20 March 2025

References

- Bashiri, F. S. et al. Object detection to assist visually impaired people: A deep neural network adventure. In *Advances in Visual Computing: 13th International Symposium, ISVC 2018, Las Vegas, NV, USA, November 19–21, 2018, Proceedings 13* (Springer, 2018).
- Fadhel, Z., Attia, H. & Ali, Y. H. Optimized and comprehensive fake review detection based on Harris Hawks optimization integrated with machine learning techniques. *J. Cybersec. Inform. Manag.* **15** (1), 1 (2025).
- Ashiq, F. et al. CNN-based object recognition and tracking system to assist visually impaired people. *IEEE Access*. **10**, 14819–14834 (2022).
- Masud, U., Saeed, T., Malaikah, H. M., Islam, F. U. & Abbas, G. Smart assistive system for visually impaired people obstruction avoidance through object detection and classification. *IEEE Access*. **10**, 13428–13441 (2022).
- Kumar, N. & Jain, A. A deep learning-based model to assist blind people in their navigation. *J. Inf. Technol. Educ. Innov. Pract.* **21**, 95–114 (2022).
- Ghasemi, Y., Jeong, H., Choi, S. H., Park, K. B. & Lee, J. Y. Deep learning-based object detection in augmented reality: A systematic review. *Comput. Ind.* **139**, 103661 (2022).
- Chen, Z. et al. IEEE. Real-time object detection, tracking, and distance and motion estimation based on deep learning: Application to smart mobility. In *2019 Eighth International Conference on Emerging Security Technologies (EST)* 1–6 (2019).
- Di Nuovo, A., Conti, D., Trubia, G., Buono, S. & Di Nuovo, S. Deep learning systems for estimating visual attention in robot-assisted therapy of children with autism and intellectual disability. *Robotics* **7** (2), 25 (2018).
- Arora, A., Grover, A., Chugh, R. & Reka, S. S. Real time multi object detection for the blind using a single-shot multibox detector. *Wirel. Pers. Commun.* **107**, 651–661 (2019).
- Pareek, K. Pixel level image fusion in moving objection detection and tracking with machine learning. *Fusion Pract. Appl.* **1**, 42–60 (2020).
- Dang, B. et al. Real-time pill identification for the visually impaired using deep learning. Preprint at <http://arXiv.org/2405.05983> (2024).
- Sharma, K. & Syal, P. Real-time object detection for assisting visually impaired people. In *2022 OPJU International Technology Conference on Emerging Technologies for Sustainable Development (OTCON)* 1–6 (IEEE, 2023).
- Alamirew, S. G. & Kebede, G. A. *Developing an Assistive Technology for Visually Impaired Persons: Ethiopian Currency Identification* (2024).

14. Yannawar, P. A novel approach for object detection using optimized convolutional neural network to assist visually impaired people. In *Proceedings of the First International Conference on Advances in Computer Vision and Artificial Intelligence Technologies (ACVAIT 2022)*, vol. 176, 187 (Springer, 2023).
15. Oluyele, S., Adeyanju, I. & Sobowale, A. Robotic assistant for object recognition using convolutional neural network. *ABUAD J. Eng. Res. Dev.* **7** (1), 1–13 (2024).
16. Triyono, L. et al. Advancing accessibility: an artificial intelligence framework for obstacle detection and navigation assistance for the visually impaired. *E3S Web Conf.* **448**, 02042 (2023).
17. Shilaskar, S. et al. Enhanced mobility for visually impaired: A smart walker with object detection. In *2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)* 1–6 (IEEE, 2024).
18. Rocha, D., Pinto, L., Machado, J., Soares, F. & Carvalho, V. Using object detection technology to identify defects in clothing for blind people. *Sensors* **23** (9), 4381 (2023).
19. Wang, M., Yao, R. & Rezaee, K. MU-Net-optLSTM: Two-stream spatial–temporal feature extraction and classification architecture for automatic monitoring of crowded Art museums. *Tsinghua Sci. Technol.* **1**, 1 (2024).
20. Mencattini, A. et al. Automatic classification of HEP-2 specimens by explainable deep learning and Jensen-Shannon reliability index. *Artif. Intell. Med.* **160**, 103030 (2025).
21. Mohanty, M. R., Mallick, P. K., Navandar, R. K., Chae, G. S. & Jagadev, A. K. Enhancing Alzheimer’s diagnosis through optimized brain lesion classification in MRI with attention-driven grid feature fusion. *Intell. Decis. Technol.* **18** (3), 1993–2018 (2024).
22. Zhou, H., Lian, Y., Li, J., Cao, X. & Ma, C. Computational spectral imaging reconstruction via a spatial–spectral cross-attention-driven network. *J. Opt. Soc. Am. A* **42** (2), 139–150 (2025).
23. Madarapu, S., Ari, S. & Mahapatra, K. A multi-resolution convolutional attention network for efficient diabetic retinopathy classification. *Comput. Electr. Eng.* **117**, 109243 (2024).
24. Cheng, C., Liu, W., Wang, X., Feng, L. & Jia, Z. DISD-Net: A dynamic interactive network with self-distillation for cross-subject multi-modal emotion recognition. *IEEE Trans. Multimedia* **1**, 1 (2025).
25. Ravinder, P. & Srinivasan, S. Automated medical image captioning with soft attention-based LSTM model utilizing YOLOv4 algorithm. *J. Comput. Sci.* **20** (1), 52–68 (2024).
26. Luan, X. et al. FMambalR: A hybrid state space model and frequency domain for image restoration. *IEEE Trans. Geosci. Remote Sens.* **1**, 1 (2025).
27. Jaiswal, T., Pandey, M. & Tripathi, P. Advancing image captioning with V16HP1365 encoder and dual self-attention network. In *Multimedia Tools and Applications* 1–25 (2024).
28. Banu, E. A. et al. Advances in computational visual information processing for neuroscience and healthcare applications. In *Integrating Machine Learning Into HPC-Based Simulations and Analytics* 329–354 (IGI Global Scientific Publishing, 2025).
29. Wu, L. et al. Deep bilateral filtering network for point-supervised semantic segmentation in remote sensing images. *IEEE Trans. Image Process.* **31**, 7419–7434 (2022).
30. Tapia, G., Allende-Cid, H., Chabert, S., Mery, D. & Salas, R. Benchmarking YOLO models for intracranial hemorrhage detection using varied CT data sources. *IEEE Access.* **1**, 1 (2024).
31. Yao, L. et al. Bearing fault diagnosis based on transfer learning with dual-flow manifold ResNet and improved CapsNet. *Meas. Sci. Technol.* **35** (7), 076123 (2024).
32. Xu, L., Teoh, S. S. & Ibrahim, H. A deep learning approach for electric motor fault diagnosis based on modified InceptionV3. *Sci. Rep.* **14** (1), 12344 (2024).
33. Ke, J. & Chen, T. Data decomposition modeling based on improved Dung beetle optimization algorithm for wind power prediction. *Data* **9** (12), 146 (2024).
34. Duan, H. & Qiao, P. Pigeon-inspired optimization: a new swarm intelligence optimizer for air robot path planning. *Int. J. Intell. Comput. Cybern.* **7** (1), 24–37 (2014).
35. <https://www.kaggle.com/datasets/thepbordin/indoor-object-detection>.
36. Heikel, E. & Espinosa-Leal, L. Indoor scene recognition via object detection and TF-IDF. *J. Imaging* **8** (8), 209 (2022).
37. Azurmendi, I., Zulueta, E., Lopez-Guede, J. M. & González, M. Simultaneous object detection and distance Estimation for indoor autonomous vehicles. *Electronics* **12** (23), 4719 (2023).
38. Mandhala, V. N., Bhattacharyya, D., Vamsi, B. & Thirupathi Rao, N. Object detection using machine learning for visually impaired people. *Int. J. Curr. Res. Rev.* **12** (20), 157–167 (2020).

Acknowledgements

The authors extend their appreciation to the King Salman center For Disability Research for funding this work through Research Group no KSRG-2024- 221.

Author contributions

Conceptualization: Abdullah M. Alashjaee, Hussah Nasser AlEisa, Abdulbasit A. Darem and Radwa Marzouk, Data curation and Formal analysis: bdullah M. Alashjaee, Hussah Nasser AlEisa, Abdulbasit A. Darem and Radwa MarzoukInvestigation and Methodology: bdullah M. Alashjaee, Hussah Nasser AlEisa, Abdulbasit A. Darem and Radwa MarzoukProject administration and Resources: Supervision; bdullah M. Alashjaee, Writing—original draft: bdullah M. Alashjaee, Hussah Nasser AlEisa, Abdulbasit A. Darem and Radwa MarzoukValidation and Visualization: bdullah M. Alashjaee, Hussah Nasser AlEisa, Abdulbasit A. Darem and Radwa MarzoukWriting—review and editing, bdullah M. Alashjaee, Hussah Nasser AlEisa, Abdulbasit A. Darem and Radwa MarzoukAll authors have read and agreed to the published version of the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to A.A.D.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025