scientific reports

Check for updates

OPEN

An intelligent object detection and classification framework for assisting visually challenged persons using deep learning and improved crow search optimization

Alaa O. Khadidos[1] & Ayman Yafoz[1,2]

According to an estimation, one billion persons are experiencing disabilities, so assistive technologies are developed, enhancing independence and accessibility. Significant developments have been made in assisting disabled people. Object detection (OD) and classification systems are effective computer technologies for image processing and computer vision (CV). It is mainly used to identify and describe objects such as vehicles, individuals, and animals from digital videos and images, which will be useful for older or disabled persons. Deep learning (DL) models demonstrate to be more expert in resolving OD defects. However, DL techniques are extensively utilized to perceive, track, and identify in real-time objects met during navigation in an indoor environment. This study proposes a Hybrid DL Model for Object Detection and Classification Using an Improved Crow Search Algorithm (HDLMODC-ICSA) method. The HDLMODC-ICSA method primarily focuses on an accurate and real-time object recognition method to assist visually challenged persons. In the initial stage, the image pre-processing stage utilizes median filtering (MF) to remove noise or distortions and make the image more transparent. Furthermore, the OD process employs the Faster R-CNN model to generate precise region proposals and detect objects within images efficiently. Moreover, the HDLMODC-ICSA technique employs the Improved LeNet-5 model to extract meaningful and discriminative features from the identified regions. The hybrid of the attention-based stacked bi-directional long short-term memory (ABS-Bi-LSTM) technique is used for OD and classification. Finally, the hyperparameter selection of the ABS-BiLSTM model is performed by implementing the improved crow search algorithm (ICSA) model. The efficiency of the HDLMODC-ICSA approach is validated by comprehensive studies using the Indoor objects detection dataset. The comparison study of the HDLMODC-ICSA approach demonstrated a superior accuracy value of 99.59% over existing techniques.

Keywords Object detection, Improved crow search algorithm, Hybrid deep learning models, Image pre-processing, Visually challenged person

Vision performs a dynamic role in recognizing the outside world, but vision loss creates more complications in everyday life. As per the world health organization (WHO), there are 285 million visually impaired people (VIPs) across the world[1]. Multiple methods are intended to assist visually challenged individuals and improve their living standards[2]. Adversely, most of these methods are restricted in their abilities. Human life depends on the five essential senses, where the ability of vision becomes the primary factor[3]. Multiple investigation analyses were developed, particularly for the design of VIP devices. Such devices were easy and sturdy, but they were lacking in usage and precision. Subsequently, such approaches have become more reliable and effective based on computers and artificial intelligence (AI)[4]. The ability to search and detect popular household objects in indoor settings, the key component of fetch-and-delivery tasks, is usually measured as one of the significant service functionalities of robots[5]. Object recognition in real scenes is a considerable difficulty in CV, as it is essential to deal with complications like occlusions, illumination variations, viewpoint changes, sensor noise, or background

[1]Department of Information Systems, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia. [2]King Salman Center for Disability Research, Riyadh 11614, Saudi Arabia. email: ayafoz@kau.edu.sa

clutter[6]. Present methods for OD are controlled by machine learning (ML) models aimed at appropriate learning demonstrations of object examples. OD is a sensible approach to making VIPs more individual[7].

It focuses on better-aiding VIPs in visualizing and navigating the outside environment. An active investigation has recently been achieved to perform intellectual methods for autonomous and safe movement VIP. These assistive models depend on AI and progressive smartphone application gadgets[8]. Object recognition is a simple process for human beings. Still, for computers, it's not a simpler task that contains a step-by-step process of identifying, recognizing, and positioning the objects with a given input degree of precision. Recognition contains detection and classification. Objects are separated into their relevant classes by performing three stages: object classification, feature extraction, and localization[9]. A multi-OD model utilizes AI and smart navigation for VIP, which have several object images highly related to the VIP and are used for training DL methods[10]. DL might be accepted to assist VIPs in OD and classification tasks, permitting them to interact and understand their surroundings more effectively. Cost-effective object classification methods are effectual devices that are incorporated with helping devices for VIPs to attain more excellent navigation in unknown settings.

This study proposes a Hybrid DL Model for Object Detection and Classification Using an Improved Crow Search Algorithm (HDLMODC-ICSA) method. The HDLMODC-ICSA method primarily focuses on an accurate and real-time object recognition method to assist visually challenged persons. In the initial stage, the image pre-processing stage utilizes median filtering (MF) to remove noise or distortions and make the image more transparent. Furthermore, the OD process employs the Faster R-CNN model to generate precise region proposals and detect objects within images efficiently. Moreover, the HDLMODC-ICSA technique employs the Improved LeNet-5 model to extract meaningful and discriminative features from the identified regions. The hybrid of the attention-based stacked bi-directional long short-term memory (ABS-Bi-LSTM) technique is used for OD and classification. Finally, the hyperparameter selection of the ABS-BiLSTM model is performed by implementing the improved crow search algorithm (ICSA) model. The efficiency of the HDLMODC-ICSA approach is validated by comprehensive studies using the Indoor objects detection dataset. The major contribution of the HDLMODC-ICSA approach is listed below.

- The HDLMODC-ICSA model employs MF to effectually remove impulse noise and improve image clarity. This pre-processing step enhances the robustness of subsequent OD and feature extraction. It confirms reliable performance even under noisy or low-quality input conditions.
- The HDLMODC-ICSA approach utilizes Faster R-CNN model to accurately detect and localize key regions of interest within the input images. This enables efficient detection of relevant features, significantly enhancing detection precision. It lays a robust groundwork for effectual downstream feature extraction.
- The HDLMODC-ICSA method implements the improved LeNet-5 architecture for extracting rich and discriminative features from the detected regions. Modifications such as deeper layers and optimized activation functions enhance feature representation. This strengthens the capability of the model to distinguish intrinsic patterns in the data.
- The HDLMODC-ICSA methodology incorporates attention-based stacked Bi-LSTM to effectually capture long-range dependencies and contextual relationships in sequential data. This is improved by the ICSA method for precise hyperparameter tuning. Altogether, they improve classification accuracy and promote better generalization across various inputs.
- The HDLMODC-ICSA approach presents a novel end-to-end pipeline that seamlessly integrates MF, Faster R-CNN, Improved LeNet-5, ABS-BiLSTM, and ICSA models. This hybrid architecture uniquely incorporates the merits of diverse techniques to improve accuracy and robustness. Its novelty is in the combined, adaptive design that ensures effectual performance on complex visual recognition tasks.

Literature of works

Abidi et al.[11] developed a solution that depends on the gradient support vector boosting-based crossover golden jackal (GSB-CGJ) method. The presented approach contains three diverse stages. The OD level is effectively implemented by applying the GSB-CGJ model. The hyper-parameters of the adaptive boosting and SVM approaches are enhanced by utilizing the golden jackal optimization model, increasing object recognition capability. Baskar et al.[12] developed a compact wearable device for VIPs that enables real-time face recognition using Multi-task Cascaded Convolutional Networks (MTCNN) methodology integrated with LAB color space processing, contrast limited adaptive histogram equalization (CLAHE), and gamma enhancement. Masal, Bhatlawande, and Shingade[13] introduced an indoor OD structure termed RSIGConv model depending on the incorporated region proposal and spatial information guided convolution. The hyper-parameters are enhanced by employing the Bayesian optimizer algorithm (BOA) to decrease train error and the difference between validation and train errors. Priyanka et al.[14] introduced a Robust OD and Tracking for VIP utilizing the Deep CNN (RODT-DCNN) method. The RODT-DCNN model includes a dual phase of operations like object classification and detection. For OD, the YOLO-v5 method was utilized. Then, the Elman neural network (ENN) method is employed for effective object classification and identification. Eventually, the sine cosine algorithm (SCA) is applied to adjust the ENN method parameter. Bai et al.[15] presented the you only look once version 8 (YOLOv8) model by integrating attention scale sequence and refining the Complete Intersection over Union (CIoU) loss with Inner-SIoU, with applying pruning and lightweight methods. Gupta et al.[16] proposed a lightweight and high-speed OD system for visually impaired users by enhancing the YOLOv6 model. This is achieved using transfer learning (TL), pruning, and finetuning techniques to improve detection accuracy and inference speed, while integrating Google Text-to-Speech (gTTs) for real-time audio feedback. Ikram et al.[17] introduced a comparative estimation of the single-shot multi-box detector (SSD) model. MobileNetv3, YOLO-v3, Faster R-CNN, and RetinaNet in real-world obstacle recognition from the camera image. In addition, the computational effectiveness regarding the time taken per frame is measured to regulate the efficiency of

every model. The workflow involves SSD application methods and image processing to identify objects like traffic signals, vehicles, and pedestrians.

Tiwary and Mahapatra[18] introduced an automated method. The presented DBN-Bald Eagle Search (DBN-BES) approach provides an effectual method for VIP that permits automated web image captions. The developed paper contains dual phases. The primary phase is image selection without captioning, and this selection process is attained by utilizing the BES technique. After the selection phase, alt text for equivalent images is created with assistance from the DBN technique. Sindhu et al.[19] developed a real-time assistive system for visually impaired individuals by utilizing Open-Source CV Library (OpenCV) for frame extraction, bootstrapping language-image pre-training (BLIP) for visual question answering (VQA), and transformer-based neural networks to interpret user voice queries and provide accurate, context-aware audio feedback. This integration of natural language processing (NLP) and CV aims to improve autonomous navigation and environmental awareness. Dang et al.[20] introduce a DL-based mobile app utilizing YOLO for real-time pill identification to assist VIPs, integrating Text-to-Speech (TTS) for auditory feedback, aiming to improve medication management and safety. Ayadi et al.[21] presented a model by generating realistic synthetic data using an optimized conditional generative adversarial network (cGAN). The model also incorporates style embedding, TL, and pre-trained convolutional neural networks (CNNs) to accurately replicate diverse Arabic calligraphic styles and improve recognition precision. Alruwaili et al.[22] introduced a real-time detection and tracking system for differently-abled individuals using DL-based YOLOv5 and YOLOv3 methodologies on RGB image datasets. Kotis, Angoura, and Lyngri[23] reviewed and analyzed the use of emerging technologies such as autonomous robots, eXtended Reality (XR), AI, digital twins (DTs), and the IoT in smart libraries to enhance accessibility and inclusivity for visually impaired people. Saini and Sengupta[24] developed a fog-cloud-based framework using an image captioning algorithm and TTS module to provide descriptive audio feedback for Blind and VIP (BVIP). Moreover, edge-sharpening and edge detection techniques enhance image clarity for users with Achromatopsia, improving their environmental understanding and independence. Sumithra, Ponnrajakumari, and Dharshini[25] proposed a secure and accessible smart ATM system for individuals with hearing, speech, and visual impairments using Quantum Cryptography, Open-Source CV (OpenCV), speech synthesis, and recognition technologies to enable independent and seamless banking transactions. Alshehri et al.[26] present a system that integrates DL and conventional image-processing techniques to translate handwritten digits into spoken language for VIPs. The system utilizes the hopfield recurrent neural network-grasshopper optimization algorithm (HRNN-GOA) methodology for digit recognition and Haralick features for improved accuracy. Qi et al.[27] introduced EmoAssist, by using large multi-modality model (LMM) and direct preference optimization (DPO) to improve empathetic responses and actionable guidance for the visually impaired community. Pongiannan, Franklin, and Richard Pravin[28] designed an app namely the BlindSpace to assist VIPs by giving scene descriptions and OD through image captioning and OD. It improves independence and confidence, providing a user-friendly, offline experience with voice commands and TTS technology. Table 1 summarizes the existing studies on OD and classification using DL and optimization techniques.

While the existing studies for VI assistive technologies exhibit great promise, various limitations still exist. Several existing models, namely GSB-CGJ, YOLO, and HRNN-GOA, are primarily focused on OD and navigation but often lack adaptability to dynamic environments or real-time interactions with complex objects. Many techniques face threats with real-world accuracy, particularly in non-controlled environments, as shown by challenges in handling lighting, textures, and movement discrepancies. Furthermore, emotional intelligence integration, as seen in EmoAssist, is still restricted in assessing human responses and giving assistance. Existing methods often overlook long-term usability, energy efficiency, and privacy concerns, which could limit widespread adoption. Additionally, models that concentrate on textual interpretation or handwriting recognition, such as HRNN-GOA, are still under-researched in their application across diverse cultural or contextual writing styles. Several models require large annotated datasets, which are often scarce or costly, restricting scalability. Real-time performance on mobile or embedded devices remains challenging due to computational constraints. Additionally, most approaches concentrate on accuracy, overlooking emotional and contextual understanding significant for VI users' holistic experience. Some methods lack robustness in diverse or complex environments, affecting reliability. User customization and adaptability to individual needs are often underexplored, reducing practical usability. There is also a gap in multimodal integration integrating audio, vision, and haptic feedback effectively. Addressing these limitations exhibits a research gap in developing lightweight, context-aware, emotionally intelligent, and adaptable systems that ensure accessibility and inclusivity for VI users in real-world settings.

Materials and methods

This study introduced an HDLMODC-ICSA method. The HDLMODC-ICSA method primarily focuses on an accurate and real-time object recognition method to aid visually challenged persons. To accomplish that, the proposed HDLMODC-ICSA method involves various methods such as image pre-processing, OD, feature extraction, classification, and hyperparameter tuning. Figure 1 depicts the overall workflow of the HDLMODC-ICSA approach.

Image Pre-processing using MF

In the initial stage, the image pre-processing stage applies MF to remove noise or distortions to make the image more transparent[29]. This model is chosen due to its robust capability to remove impulse (salt-and-pepper) noise while preserving crucial edge details. Unlike mean or Gaussian filters that blurrs image features, MF maintains structural integrity, which is significant for accurate OD and feature extraction. Its non-linear nature allows it for handling high-frequency noise efficiently without altering the underlying image content. MF is also computationally effectual and easy to implement, making it appropriate for real-time or large-scale applications.

Authors	Techniques	Metrics and Dataset	Performance, Strengths and Limitations
Abidi et al[11].	GSB-CGJ, adaptive boosting, SVM	Accuracy, Precision, Recall, F1-Score, AUC, Execution Time, Image and Video Dataset, Intel RealSense Camera	High accuracy and real-time performance; robust adaptability and reliability; mitigated execution time; limited details on scalability.
Baskar et al[12].	Face Recognition, MTCNN, LAB Color Space, CLAHE, Gamma Enhancement	CPU Usage, Memory Usage, Frames Per Second (FPS), Model Load, Average CPU Load, Real-time Embedded Data	Efficient real-time face recognition; compact and portable; limited to specific hardware constraints.
Masal, Bhatlawande, and Shingade[13]	RSIGConv, BOA, Spatial Information Guided Convolution	Accuracy, Train Error, Validation Error, SUN RGB-D	High accuracy (97.77%) with effective feature fusion; may need complex computations for real-time use.
Priyanka et al[14].	RODT-DCNN, YOLO-v5, ENN, SCA	OD and Classification Accuracy, Parameter Optimization, Standard OD Dataset	Improved detection with optimized parameters; complexity may affect real-time performance.
Bai et al[15].	YOLOv8, Attention Scale Sequence, Inner-SIoU Loss Function, Pruning and Lightweight Methods	Precision, Recall, AP, mAP, FLIR and WOTR Dataset	Improved detection accuracy and efficiency with significant model compression.
Gupta et al[16].	TL, YOLOv6 Baseline, Pruning & Finetuning, gTTs	AP, FPS, MS-COCO Dataset	High detection accuracy and fast inference speed; efficient for real-time embedded use.
Ikram et al[17].	YOLOv3, MobileNetv3, RetinaNet, Faster R-CNN, Image augmentation	Precision, Recall, F1 Score, IoU Thresholds, Time Per Frame, Camera Images (Real-Time)	YOLOv3 exhibited top precision (96%) with robust real-time efficiency.
Tiwary and Mahapatra[18]	DBN-BES, Automatic captioning	Accuracy, Caption Relevance, Image-Text Matching, Web Image Dataset, Uncaptioned Image Data	DBN-BES enhances image captioning and accessibility but face difficulty with complex images.
Sindhu et al[19].	OpenCV Frame Extraction, Voice-to-text Conversion, BLIP VQA, Transformer Networks	Response Accuracy, Processing Time, Audio Feedback Latency, Visual Question Answering (VQA), Real-Time User Videos	Scene understanding is improved with minor delays in complex or dynamic cases.
Dang et al[20].	YOLO, DL, TTS	Detection Accuracy, User Experience, Pill Image Dataset, Real-Time Mobile Data	Accurately detects pills with real-time feedback, though performance may vary with pill similarity or poor lighting.
Ayadi et al[21].	Optimized GAN, CGAN, Style embedding, TL, Pre-trained CNNs	Accuracy, Recall, F-Score, Handwritten Arabic dataset	Achieves 99% accuracy with high-quality synthesis.
Alruwaili et al[22].	YOLOv5, YOLOv3	Precision, Recall, mAP, RGB images dataset, 17,079 PNG files, 15,278 YML annotations, 5 classes (pedestrian, wheelchair user, etc.)	YOLOv3 outperforms YOLOv5 in detection accuracy but may have higher computational cost.
Kotis, Angoura, and Lyngri[23]	Autonomous Robots, XR, AI, DT, IoT	Accessibility Assessment, User Engagement, Inclusivity Measures, Smart Library User Data, Visually Impaired Interaction Logs	Provides inclusive smart library solutions but encounters difficulty with diverse accessibility.
Saini and Sengupta[24]	Fog-cloud framework, Image captioning algorithm, TTS, Edge-sharpening and detection technique	Accuracy (%), Performance comparison, Custom BVIP image dataset (implied)	Attains 96.84% accuracy with clear descriptions and edge enhancement; may face difficulty in complex settings.
Sumithra, Ponnrajakumari, and Dharshini[25]	Quantum cryptography, OpenCV, Speech synthesis and recognition, Audio feedback, Haptic interfaces	Transaction security, User accessibility, Interface usability, Banking transaction data, Speech command datasets	Ensures secure, accessible banking for disabled users but has complexity in quantum tech implementation.
Alshehri et al[26].	HRNN-GOA, Haralick features extraction, DL, Image processing	Accuracy, Precision, Recall, Specificity, F1-Score, AUC, False Positive Rate, Handwritten Digit Datasets	Accurate digit recognition; requires adaptation for complex characters.
Qi et al[27].	LMMs, DPO, Emotion-Assistive Model	Empathy and Suggestion Score, Emotional Recognition, EmoAssist Benchmark	Improves emotional support for visually impaired; requires real-world testing.
Pongiannan, Franklin, and Richard Pravin[28]	Image Captioning, OD, Voice Commands, TTS	Description Accuracy, Detection Precision, User Satisfaction, Custom Captured Images, Public Object Datasets	Accurate offline scene descriptions and OD; encounter challenges with complex or cluttered environments.

Table 1. Summary of utilized techniques, evaluation metrics, datasets, and key performance of various assistive technologies for visually impaired users.

Compared to more intrinsic denoising techniques, MF presents a balanced trade-off between performance and simplicity. This makes it an ideal choice for improved image quality in the proposed pipeline. Figure 2 specifies the structure of the MF model.

MF is a non-linear image processing mode employed to decrease noise, particularly salt-and-pepper noise while maintaining edges in an image. In object recognition and classification tasks, it aids in improving image quality by eliminating unwanted noise, which can inhibit precise object localization and feature extraction. MF smoothing an image creates more effective edge recognition and segmentation models. It is mainly beneficial in lower-quality or noisy images, enhancing the performance of conventional and DL-based methods. However, extreme filtering can blur significant features, so it must be used sensibly. Overall, MF is a beneficial pre-processing stage for enhancing classification and detection accuracy.

OD using faster R-CNN method

Next, the OD process is performed by the Faster R-CNN model to generate precise region proposals and detect objects within images efficiently[30]. This model is chosen due to its high accuracy and efficiency in detecting objects with varying scales and complex backgrounds. Unlike conventional methods or earlier R-CNN variants, it combines region proposal and classification into a single, end-to-end trainable network, significantly mitigating processing time. Its region proposal network (RPN) allows for fast and precise localization of objects, which is crucial for downstream tasks like feature extraction. Compared to models such as YOLO or SSD, Faster R-CNN gives superior detection performance, particularly in conditions needing high localization precision.

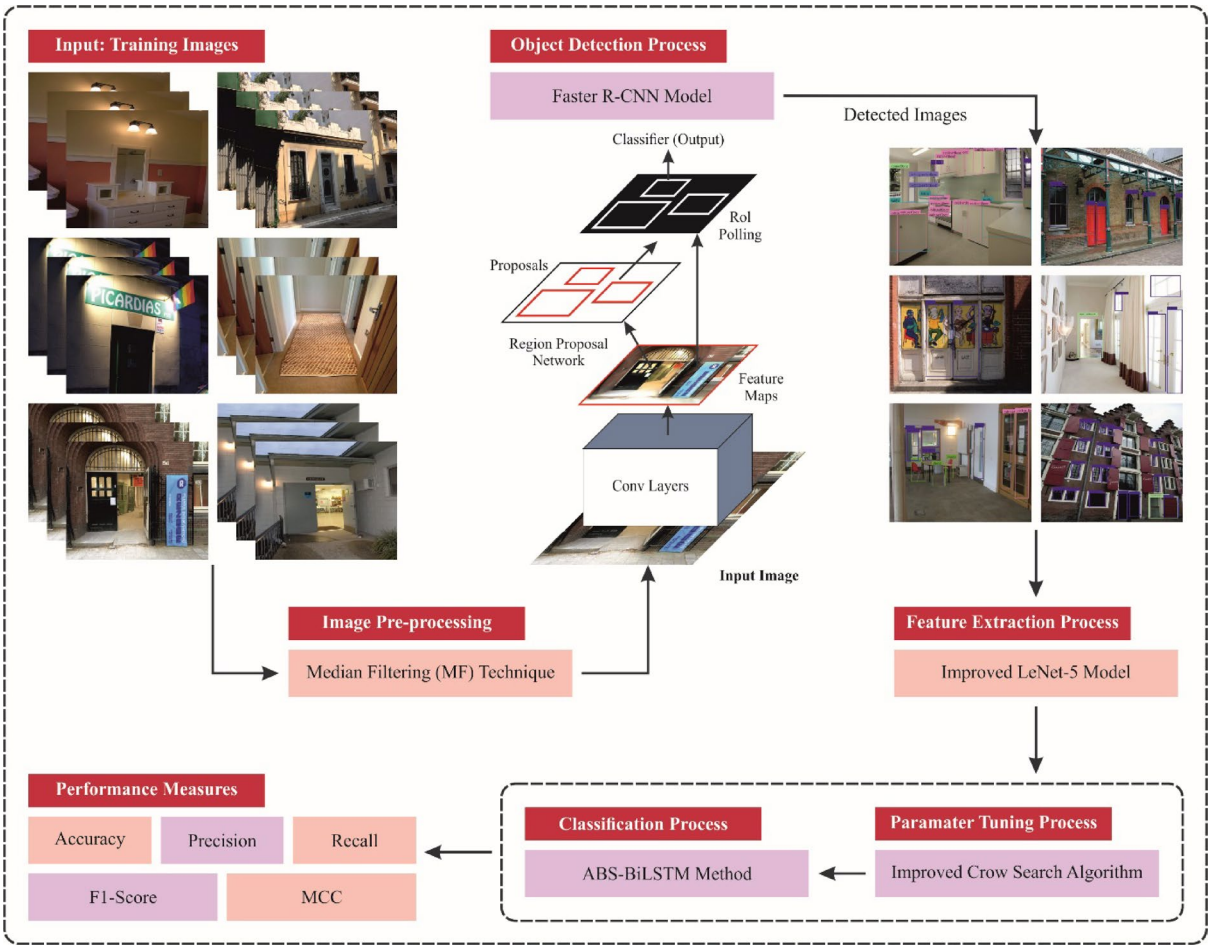

Fig. 1. Overall workflow of HDLMODC-ICSA approach.

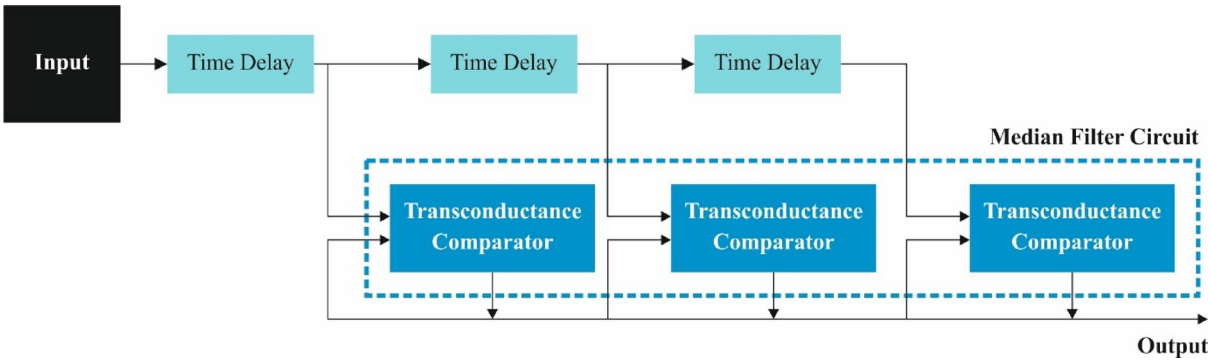

Fig. 2. MF architecture.

Its robustness across diverse datasets makes it a reliable choice for real-world applications. This justifies its role as the OD backbone in the proposed model. Figure 3 demonstrates the structure of the Faster R-CNN method.

Faster R-CNN is the renowned OD model for its higher precision. It contains dual modules: a detection system and RPN. The key basis of Faster R-CNN includes the succeeding stages:

1) base network: utilize a pre-trained CNN to remove features from the input image
2) RPN: It makes candidate targeted areas by sliding windows over the feature mapping, considering them as anchor boxes, and forecasting whether the anchor boxes comprise objects and how to fine-tune the limits of the anchor boxes.

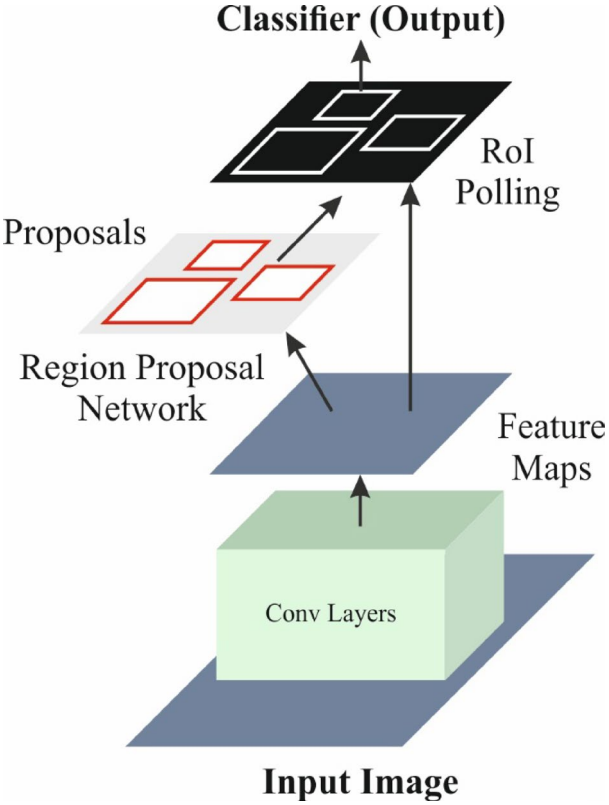

Fig. 3. Faster R-CNN framework.

3) Region of interest pooling (RoI Pooling): Split all candidates' areas into predetermined size subfields and map them on the static size feature maps.
4) Object classification networks: Utilize fully connected (FC) systems for classifying all candidate areas, seizing the RoI Pooling output as input, and outputting the possibilities of all candidate areas characterized by various targeted classes.
5) Bounding box regression: Implement this regression to fine-tune the coordinates of the bounding box targets in all candidate areas.

To carry out OD in Faster R-CNN, they utilize the RPN to make candidate areas. The RPN phase includes dual primary equations, one to compute the coordinate of anchor boxes and the other to calculate the loss amongst ground truth bounding boxes and the anchor boxes. At last, they utilize the succeeding equations for computing the anchor boxes coordinates and establish their locations within the images:

$$\begin{aligned} x_{anchor} &= x_{center} - \frac{w_{anchor}}{2} \\ y_{anchor} &= y_{center} - \frac{h_{anchor}}{2} \\ w_{anchor} &= width_{anchor} \\ h_{anchor} &= height_{anchor} \end{aligned} \tag{1}$$

x_{anchor} and y_{anchor}: The upper left coordinate of the anchor boxes represents their location within the images. x_{center} and y_{center}: The coordinates of the central point of the anchor or target boxes are applied to define the location of the anchor boxes. w_{anchor} and h_{anchor}: The width and height of the anchor boxes utilized to establish their dimensions. $width_{anchor}$ and $height_{anchor}$: The pre-defined width and height of the anchor boxes, normally fixed as static values in training. Continuing these, they present the Smooth Ll Loss as the loss function amongst ground truth bounding and anchor boxes. The calculation equation for these loss functions is demonstrated:

$$L_{bbox} = \sum\nolimits_{i} L_{smooth}\left(t_i - t_i',\ 1_i\ is\ positive\right) \tag{2}$$

Whereas t_i characterizes the forecast bounding boxes offset, t_i' characterizes the consistent ground truth bounding boxes offset, and 1_i means positive is the indicator function, which captures a value of 1 after the

anchor boxes i represent positive instances. During Faster $R-CNN$, there is additionally an essential phase of computing the output of the RoI pooling layer. The pooling layer of RoI is applied for mapping RoIs of dissimilar dimensions on a predetermined size feature mapping to maintain the RoI feature's spatial alignment. The succeeding equation is utilized to calculate the RoI pooling layer output:

$$F_{roi} = RoIpooling\left(F_{conv}, p\right) \tag{3}$$

F_{conv} characterizes the convolution feature mapping gained from the feature extraction system, and p symbolizes the input parameters for the pooling layer of RoI, using the size information and coordinates of the RoI. The pooling layer of RoI maps RoI's of various sizes on predetermined size feature maps to maintain the RoI feature's spatial alignment.

The principle of this stage is to carry out pooling processes on the feature mapping areas comparable to $RoIs$ of dissimilar dimensions, leading to predetermined size RoI features. This permits mapping $RoIs$ of dissimilar dimensions on the similar-sized feature mapping, enabling succeeding bounding box regression and object classification. The detection system captures the candidate boxes from the RPN as input and implements bounding box regression and object classification. They utilize the following formulations to calculate the detection system outputs. The equation for object classification is as shown:

$$F_{cls} = softmax\left(W_{cls} \cdot F_{roi} + b_{cls}\right) \tag{4}$$

The equation for bounding box regression is as shown:

$$F_{reg} = W_{reg} \cdot F_{roi} + b_{reg} \tag{5}$$

Whereas $W_{cls}, b_{cls}, W_{reg}$, and b_{reg} are learned parameters. The softmax function is applied to transform the output of object classification into the likelihood distribution above class labels. Finally, they utilize the bounding box regression and object classification outcomes to filter out the end recognition outcomes. By establishing a threshold, they choose the targeted boxes with higher confidence as the last recognition outcomes and enhance their bounding box locations utilizing the bounding box regression outcomes for more precise localization. By incorporating the detection network and RPN, Faster R-CNN attains an accurate OD level and has proven essential performance developments on numerous benchmark datasets.

Feature extraction using improved LeNet-5 method

Furthermore, the proposed HDLMODC-ICSA method employs the Improved LeNet-5 model to extract meaningful and discriminative features from the identified regions[31]. This model is chosen due to its simplicity, efficiency, and adaptability to modifications that improve performance. By extending the original architecture with deeper layers, advanced activation functions, or batch normalization, the model can capture more complex and discriminative features. Compared to heavier networks namely VGG or ResNet, the improved LeNet-5 presents a lightweight alternative with lesser computational cost while maintaining robust feature representation. It is specifically efficient for mid-level feature extraction from localized regions, as given by Faster R-CNN. Its fast training and reduced parameter count make it ideal for integration in hybrid pipelines. These merits make it a balanced and effectual choice for the proposed framework. Figure 4 demonstrates the Improved LeNet-5 approach.

LeNet-5 method is presented as a CNN base method. The approach contains a 7-layer structure and can execute pooling, convolution, and incorporating information functions. LeNet-5 model architecture in which C represents the convolutional layer, S signifies the pooling layer, and F^1 indicates an FC layer. All neurons on layer Cl are linked to an input neighbourhood. Layer S2 is a pooling layer gained by sampling from layer Cl. All neurons in layer F6 are connected to each of the neurons in the preceding layer. Convolutional processes improve the novel signal features and decrease noise.

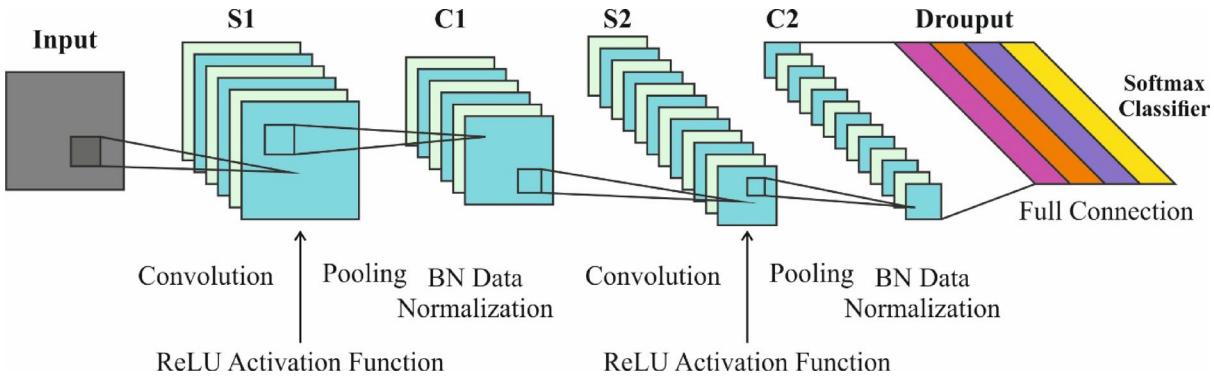

Fig. 4. Structure of Improved LeNet-5 model.

$$x_j^l = f\left(\sum_{i\in M_j} x_i^{l-1} W_{ij}^l + b_j^l\right) \tag{6}$$

Whereas l characterizes the layer, W signifies the convolutional kernel weighting parameter, M_j symbolizes a selection of input features, and b denotes the bias.

The combination of feature categories is carried out throughout the down-sampling process to decrease the spatial sizes and, therefore, mitigate the occurrence of overfitting. When the input feature count is n, the feature counts after the down-sampling layer might be lower than or equivalent to n, and the output feature dimensions should be smaller.

$$x_j^l = f\ \left(down\ \left(x_j^{l-1}\right) + b_j^l\right) \tag{7}$$

During this FC process, a single neuron in the FC layer is linked to each of the neurons in the previous layer. The FC layer incorporates the class-distinctive local information from the pooling and convolutional layers and transforms the feature information into a 1D vector. At last, the output of the FC layer is given to the output layer. To model the reduction of calculation and training time, LeNet-5 CNN is enhanced in this study.

Initially, a small serial convolutional kernel was applied to replace the convolutional kernel at layer C3 within the LeNet-5 model. Dual 3×3 size convolutional kernels substitute the 5×5 size convolutional kernel. It is discovered that the parameter counts for convolutional cost utilizing dual 3×3 convolutional kernels are lower than one 5×5 convolutional kernel. Simultaneously, the network layer counts improve after a small convolution kernel is applied rather than the unique convolution kernel. The rise in the network layer counts assists in enhancing the precision of the model classification. Adding an activation function after the added convolutional layers enhances the method's non-linear ability and enables the fitting of more composite functions. The probability of a convolutional layer regarding a single hidden neuron is:

$$P\,(v,\ h) = e^{-E(v,h)} \tag{8}$$

On the other hand, $E(v,\ h)$ characterizes the energy function under the Bernoulli distribution.

$$E\,(v,\ h) = -\sum\nolimits_{j=1}^{m} b_j v_j - \sum\nolimits_{i=1}^{n} c_i h_i - \sum\nolimits_{i=1}^{n}\sum\nolimits_{j=1}^{m} w_{ij} v_j h_i \tag{9}$$

The probability-based solution reflects only 1 neuron; formerly, the probability of n hidden neurons presenting on the convolutional layer is stated as:

$$P\,(v) = \sum\nolimits_{h} P\,(v,\ h) = \sum\nolimits_{h} e^{-E(v,h)} \tag{10}$$

The probability of m neurons in the convolutional layer over the hidden layer (HL) is stated as shown:

$$P\,(h) = \sum\nolimits_{v} P\,(v,\ h) = \sum\nolimits_{v} e^{-E(v,h)} \tag{11}$$

The probability that the ith hidden neuron is activated for m visible neurons is demonstrated as:

$$P\,(h_i = 1|v) = \sigma\ \left(c_i + \sum\nolimits_{j=1}^{m} w_{ij} v_j\right) \tag{12}$$

Whereas w denotes the weight between the dual layers, c and b signify the respective offsets. For n hidden neurons, the probability that the jth visual unit is activated is indicated as shown:

$$P\,(v_j = 1|h) = \sigma\ \left(b_j + \sum\nolimits_{i=1}^{n} w_{ji} v_i\right) \tag{13}$$

Here, $\sigma\ (\bullet\,)$ signifies the function of machine probability.

$$\sigma\ (x) = sigmoid\,(x) = \frac{1}{1+e^{-x}} \tag{14}$$

Formerly, the activation function of $ReLU$ is applied to replace the activation function of Sigmoid. During this similar training atmosphere, the activation function of $ReLU$ needs less training than the Sigmoid function. For N input samples $v = \{v_0,\ v_1,\ \ldots\ ,\ v_N\}$, and $v_0,\ v_1, \ldots\ , v_N$ follow independent distributions.

$$P\,(v) = \prod\nolimits_{t=0}^{N} P\,(v_t) \tag{15}$$

The probability estimation for the sample set v is said as follows:

$$L\,(\theta\,) = \prod\nolimits_{t=0}^{N} P\,(v_t|\theta\,) \tag{16}$$

Here $\theta = \{w,\ c,\ b\}$ signifies the energy parameter.

Solving for the maximal value of $L(\theta)$ converts into finding a solution for the maximum value of $L(\theta)$.

$$\widehat{\theta} = \text{argmax}_{\theta} L(\theta) = \text{argmax}_{\theta} \sum_{t=0}^{N} \ln P(v_t|\theta) \tag{17}$$

$$\theta^{*} = \theta + \eta \frac{\partial \ln P(v)}{\partial \theta} \tag{18}$$

Now, η refers to the learning rate, and $\eta > 0$..

The sample is logarithmically resolved for a single sample $v_0 = \{v_{01},\ v_{02},\ \ldots,\ v_{0m}\}$.

$$\ln P(v_0) = \ln \sum_{h} e^{-E(v_0,h)} - \ln \sum_{v,h} e^{-E(v,h)} \tag{19}$$

Employ the partial derivative to $\theta = \{w,\ c,\ b\}$.

$$\frac{\partial\ \ln P(v_0)}{\partial \theta} = -\sum_{h} P(h|v_0) \frac{\partial E(v_0,h)}{\partial \theta} + \sum_{v,h} P(v,\ h) \frac{\partial E(v,h)}{\partial \theta} \tag{20}$$

.

The probability distribution is exposed as shown:

$$P(v,\ h) = P(h|v) P(v) \tag{21}$$

The partial derivatives of $\theta = \{w,\ c,\ b\}$ are gained for all three parameters w, c, b.

$$\frac{\partial \ln P(v_0)}{\partial w_{ij}} = P(h_i = 1|v_0) v_{0j} - \sum_{v} P(v) P(h_i = 1|v) v_j \tag{22}$$

$$\frac{\partial \ln P(v_0)}{\partial b_j} = v_{0j} - \sum_{v} P(v) v_j \tag{23}$$

$$\frac{\partial\ \ln P(v_0)}{\partial c_i} = P(h_i = 1|v_0) - \sum_{v} P(v) P(h_i = 1|v) \tag{24}$$

.

Then, the S4 pooling layer of the LeNet-5 CNN is adjusted with the spatial pyramid pooling (SPP) model to reduce the influence of the pooling process on the feature values. The cubes characterize the output of the feature maps from the convolutional layers. These feature maps are connected to the three pooling layers to obtain size 4×4, 2×2, and 1×1 outputs. The outputs of the three pooling layers are connected to get a 21-dimensional vector. The results of performance are compared with the method with and without SPP. After assuming the SPP model, the error rate of the single-size and multiple-size trained network methods was reduced by 0.62% and 1.12%, respectively. Lastly, an FC layer is applied to replace the C5 layer of the LeNet-5CNN. After the development, three convolutional layers are earlier than the C5 network layer. With no reduction in the convolutional layers, substituting the C5 layer with an FC layer additionally increases the stability and classification precision of the method.

Classification using ABS-BiLSTM model

For OD and classification, the hybrid of the ABS-BiLSTM technique is employed[32]. This model is chosen due to its capability to capture both past and future context in sequential data, which is significant for comprehending intrinsic patterns. The bidirectional structure improves the capability of the method to learn temporal dependencies, while the attention mechanism concentrates on the most relevant features, enhancing interpretability and accuracy. Compared to standard LSTM or unidirectional models, this technique exhibits superior performance in tasks needing deep contextual understanding. It is particularly effectual in scenarios with variable-length input and subtle feature variations. Its incorporation confirms robust classification outcomes in the proposed system. These advantages make it an ideal choice for high-level decision-making tasks. Figure 5 represents the structure of the ABS-BiLSTM model.

A stacked bi-directional LSTM neural network (NN) is the attention-based method fundamental to the architecture. After transforming the text into a fixed-length vector contribution, the structure is initially used to clean the data of the text, remove specific capitalization and characters, and keep only the particulars, which are retrieved as semantics. To discover the syntax of the sentence here, the features of the sequence of the text are initially examined by utilizing stacked Bi-LSTM. Then, the self-attention (SA) layer powerfully weighs characteristics, emphasizing contextual significance and tackling ambiguous words. Finally, a multi-layer view is applied to make the classification results. Embedding layer: By embedding usage, words are currently identified by computers and transformed into vectors. The NN language method allows NN and probability-based language processing for text methods by defining probability-based language method features through NNs. After a sentence is lengthy, $h\{C_1,\ C_2,\ \ldots\ C_i,\ \ldots\ C_h\}$, then utilize the learned word vector modelling to perform word vector mapping, which makes a word vector matrix from the sequence of words

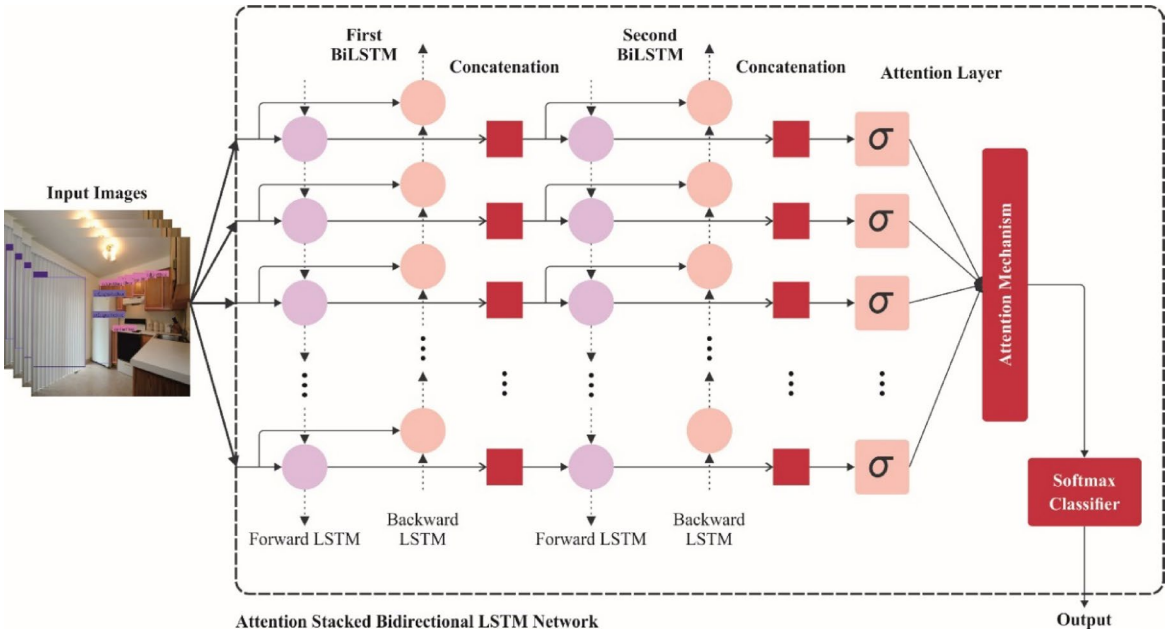

Fig. 5. Structure of ABS-BiLSTM model.

$\{D_1, D_2, \ldots D_j, \ldots D_t\}$, while the size of the matrix is $h*d$, in such cases, d denotes the size of the word vectors.

Stacked Bi-LSTM-NN layer: An individual NN could efficiently remove complex data features. Deeper frameworks, namely the attention-based stacked Bi-LSTM method, utilize the initial Bi-LSTM layer output as input for the next. These models more efficiently solve complex things, thus improving the feature representation abilities of the model. Primarily the layer of Bi-LSTM $\{D_1, D_2, \ldots D_j, \ldots D_t\}$, HL is typically transferred by the initial layer of Bi-LSTM, while i represents an i^{th} word, the parameters C_1, z_1 be spread to the members of the layers. The connection of the forward and the reversed LSTM's HL gives the information for the following layer of Bi-LSTM, helping as the initial Bi-LSTM layer output. The features vector representation is made through the next layer of Bi-LSTM t_i of text by uniting HLs, which are forward and reverse. t_i refers to a graph that contains implied deeper-level relations in a text that are important to increase the classification precision. The succeeding Eqs characterize the HL ti1. (25) −(27):

$$\overrightarrow{t_i^1} = \sigma\left(C_1\left[\overrightarrow{t_{i-1}^1, d_i}\right] + z_1\right) \tag{25}$$

$$\overrightarrow{t_i^2} = \sigma\left(C_1\left[\overrightarrow{t_{i-1}^2, d_i}\right] + z_1\right) \tag{26}$$

$$t_i^1 = \overrightarrow{t_i^1} \oplus \overleftarrow{t_i^1} \tag{27}$$

The following equations define the stacked Bi-LSTM's final output within the next layer Eqs. (28) −(30):

$$\overrightarrow{t_i^2} = \sigma\left(C_2\left[\overrightarrow{t_{i-1}^2, t_i^1}\right] + z_2\right) \tag{28}$$

$$\overrightarrow{t_i^2} = \sigma\left(C_2\left[\overrightarrow{t_{i-1}^2, t_i^1}\right] + z_2\right) \tag{29}$$

$$t_i^2 = \overrightarrow{t_i^2} \oplus \overleftarrow{t_i^2} \tag{30}$$

Attention layer: Afterward, the stacked layer of Bi-LSTM, an SA mechanism, is used for weighting all contextual representations of word vectors, making the attention layer imitate every word's importance to sentence semantics. Lastly, the proportionate summary is applied to make the global semantical representation of the sentence S.

Stage1: Calculate the weighted score e_i that computes how much ith word gives to the understanding of uncertain terms. Equation (31)' s output h_j through the FC technique having an activation objective of $tanh$ acts as a computing process.

$$e_i = tanh\left(C_c h_j + z_c\right) \tag{31}$$

The computation of the activation function of $tanh$ is Eq. (32):

$$tanh\,(d) = \frac{u^{d} - u^{-d}}{u^{d} + u^{-d}} \tag{32}$$

Whereas h_i denotes a layer that particularly hides the adequate for the C_c training parameter and the z_c terms for bias denotes a state, ith expression is applied to define a hyperbolic at a tangent function, which captures some actual integer as input and outputs a value between $(0, 1)$. It is applied to process information and outputs; the nearer the output is to 1.0.

Stage 2: Define the attention weight of the word $\alpha_{\,i}$. The function of Softmax is applied next standardization e_i^T and u_c dot-creation exposed in Eq. (33):

$$\alpha_{\,j} = \frac{\exp\left(e_i^T e_c\right)}{\sum_{i=0}^{t} \exp\left(e_i^T e_c\right)} \tag{33}$$

Here, e_c denotes the learnable context vector by an arbitrary beginning.

Stage 3: Eq. (34) displays how the computed weights for attention are used to stack Bi-LSTM's HL outputs or clause vector S.

$$S = \sum_{i=0}^{t} \alpha_{\,i} h_i \tag{34}$$

The representation vector might be calculated using the model mentioned above for phrases with an SA mechanism.

The layer of output: The focal layer outputs a higher-dimensional depiction, S, which is used as the feature vector. This feature vector, therefore, connects to the FC-HL, whereas the activation function of softmax, as defined in Eq. (35), transforms it into the $N - dimensional$ vectors.

$$\widehat{a} = Softmax\,(MLP\,(S)) \tag{35}$$

For training the projecting method, calculate the loss function of cross-entropy (P) between the ground-truth label a_j and an anticipated label $\widehat{a}$ utilizing Eq. (36) as shown:

$$P\,(\theta\,) = -\sum_{a_j \in\, a, \widehat{a} \in\, \widehat{a}} \widehat{a}_i \log \widehat{a}_i + (1 - a_i) \log\,(1 - \widehat{a}_i) \tag{36}$$

θ signifies the ability of the training parameter vector. Adjusting a parameter, which reflects the varying slope of the loss process, $\theta\,.\nabla_{\,\theta}\,P\,(\theta\,)$ reduces the cross-entropy loss function utilizing a gradient descent optimizer method θ. Equation (37) denotes the value of the gradient descent model updated equation.

$$\theta\ = \gamma\,.\nabla_{\,\theta}\,P\,(\theta\,) \tag{37}$$

γ displays the learning rate.

Hyperparameter tuning using ICSA model

Eventually, the ICSA design will implement the hyperparameter selection of the ABS-BiLSTM model. This technique is chosen due to its effectual global search capability, inspired by the natural behavior of crows. ICSA outperforms in exploring the solution space with lesser iterations and averting local optima, which is significant for finding the optimal combination of hyperparameters. Unlike gradient-based methods, ICSA does not depend on differentiability and can efficiently tune non-differentiable or complex objective functions. Its capability to balance exploration and exploitation confirms robust optimization without excessive computational cost. Furthermore, the flexibility of the ICSA model makes it appropriate for fine-tuning the hyperparameters of DL methods, giving significant enhancements in performance and generalization. This makes ICSA an ideal choice for fine-tuning the proposed model. Figure 6 indicates the flowchart of the ICSA model.

The CSA is the population-based intelligent optimizer model stimulated by the crow's foraging behaviour[33]. By mimicking crows' defence and memory tactics, the CSA seeks out optimum solutions inside the solution area. It is considered because of its extensive applicability and simplicity, having established efficient outcomes through different optimizer difficulties. Nevertheless, the CSA tends to get stuck in local bests in the searching procedure, and its balance between exploitation and exploration is narrow, resulting in slow convergence speeds and inadequate accuracy in multimodal, complex, higher-dimension issues. To deal with these restrictions, this work presents an ICSA that incorporates reinforcement learning (RL) strategies and adaptive neighbourhood search. Initially, adaptive neighbourhood search powerfully fine-tunes the crow's search radius, guaranteeing exploration flexibility throughout dissimilar searching stages; this permits wide-ranging searching in the initial phases while slowly concentrating on adjustment in the end phases. Additionally, the combination of RL permits the crows to dynamically choose optimum searching behaviours depending on the response from the atmosphere, thus improving the model's decision-making abilities. This development supports the CSA's local searching precision and global exploration abilities and considerably improves convergence speed and solution accuracy, improving stability and robustness in challenging composite optimizer issues. The complete stages and main expressions for the ICSA are as demonstrated:

(1) Population Initialization: Initializing the location x_i and memory location m_i of all crows within the population, in addition to their respective fitness values. Set the adjustable neighbourhood radius $\delta_{\,j}$ and rate

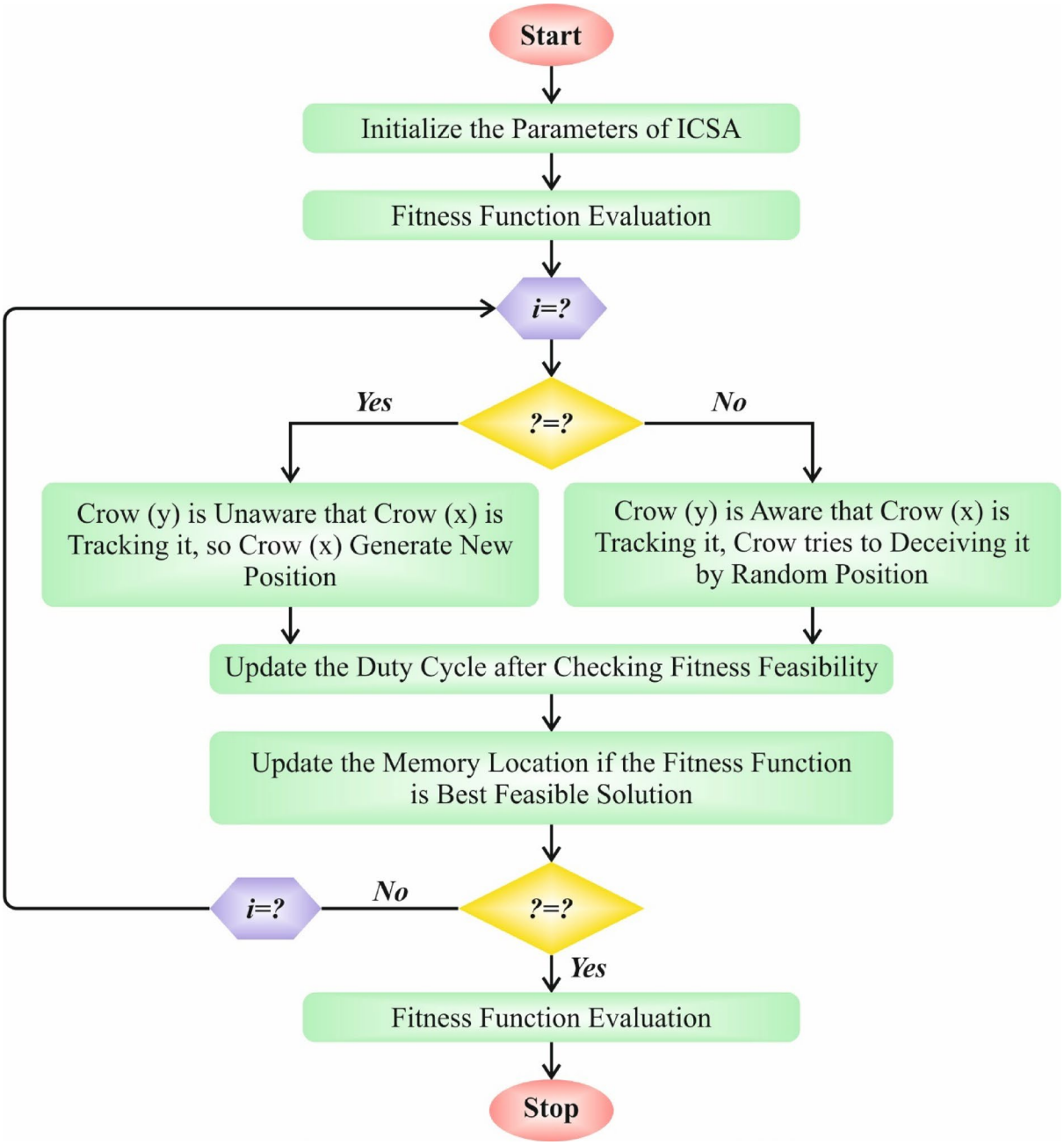

Fig. 6. Flowchart of ICSA methodology.

of learning α_i, along with the Q-table for RL, which should be applied to learning the optimum behaviours for various conditions.

(2) Adaptive Neighborhood Search: Establishing the neighbourhood radius δ_i and dynamically fine-tuning the searching neighbourhoods. Describe the primary radius. δ_{init} and scaling features γ (whereas $\gamma \in (0,1)$) Fine-tune the neighbourhood radius according to the present population fitness and density, permitting the searching variety to be dynamically modified in the optimizer procedure (38):

$$\delta_i = \delta_{init} \times \gamma^{\frac{1}{t+1}} \tag{38}$$

Whereas t denotes the present iteration, and γ controls the rate of contraction.

(3) RL Approach: Applying Q-learning in RL to determine a behaviour choice approach. Describe the state S (with information such as present fitness and location) and the actions A (selecting extended or neighbourhood searches). Updated all crow's Q-value through the rate of learning η and factor of discount λ. The Q-value updated equation is presented in Eq. (39):

Input:

- Number of crows N
- Maximum iterations T
- Search space boundaries
- Initial neighborhood radius δ_{init}
- Initial learning rate α_{inii}

Output:

- Best solution found X_{best}

Step 1: Initialization

1.1. Initialize the position X_i and memory M_i for each crow $i = 1,2,\ldots,N$ arbitrarily within the search space.

1.2. Initialize the fitness value for every crow.

1.3. Set the neighborhood radius $\delta_i = \delta_{init}$ and learning rate $\alpha_i = \alpha_{inii}$.

1.4. Initialize Q-tables for each crow for RL-based behavior selection.

Step 2: Adaptive Neighborhood Radius

2.1. For each iteration t, update δ_i to gradually mitigate the search radius:

- Large radius in early iterations for exploration
- Smaller radius in later iterations for exploitation

Step 3: RL for Behavior Selection

3.1. Determine the current state based on fitness and position for every crow.

3.2. Choose an action using the Q-table (local or global search)

3.3. Use the outcome (reward received) to update the Q-value

Step 4: Position Update

4.1. Move the crow toward another crow's memory using a wider neighborhood range; if the chosen action is global search.

4.2. Move the crow toward its own memory using a narrower step; if the chosen action is local search.

4.3. Add stochastic variation to encourage exploration.

Step 5: Memory Update

5.1. Update M_i with the new position if it has better fitness than the stored memory.

Step 6: Adaptive Parameter Update

6.1. For balancing exploration and exploitation dynamically, adjust α_i and δ_i based on population diversity and iteration count

Step 7: Termination

7.1. Terminate the algorithm, if the stopping criterion is met (maximum iterations or convergence).

7.2. Return the best solution found X_{best} from all crow memories.

Algorithm 1: ICSA technique.

$$Q(S,A) \leftarrow Q(S,A) + \eta \left[r + \lambda \cdot \max_{A'} Q\left(S', A'\right) A' - Q(S,A) \right] \tag{39}$$

Here r denotes the reward gained from the present action, and S' denotes a novel state deriving from action A..

(4) Updated Position: According to neighbourhood search and RL decisions, select whether to use the memory location (for local search) or implement global exploration in Eq. (40):

$$X_i^{t+1=} \begin{cases} x_i^t + r_i \cdot \delta_j \cdot (m_j - x_i^t), & if\ global \\ x_i^t + r_i \cdot \alpha_i \cdot \left(m_i - x_i^t\right), & if\ local \end{cases} \tag{40}$$

Now, r_i denotes a randomly generated number between $0\ and\ 1$.

(5) Updates Memory Location: When the novel location X_i^{t+1} is superior to the memory location m_i, updated m_i through X_i^{t+1}..

(6) Adaptive Modification of Learning Rate and Neighborhood Radius: According to the crow's population diversity and fitness, adaptively fine-tune the learning rate α_j and neighbourhood radius δ_i to guarantee a more excellent exploration range in the initial phases and more concentrated local searches later in Eq. (41):

$$\begin{aligned} \delta_i &= \delta_{init} \times \exp\left(-\beta \cdot \tfrac{t}{T}\right) \\ \alpha_i &= \alpha_{inii} \times \exp\left(-\gamma \cdot \tfrac{t}{T}\right) \end{aligned} \tag{41}$$

Here, γ and β control the decay rates of the learning rate and radius, correspondingly, and T denotes maximal iteration counts.

(7) Stopping Conditions: The iteration procedure stops when a pre-defined maximal iteration count is reached or when convergence conditions are encountered (for example, the fitness value displays no vital development), and the optimum solution is produced. Algorithm 1 represents the ICSA model.

Table 2 Specifies the hyperparameter settings for the ICSA model. Each parameter is defined along with its symbol, functional role within the optimization process, and general value ranges based on standard experimental setups. These parameters are significant for balancing exploration and exploitation, controlling adaptive search behavior, and ensuring robust convergence in complex optimization environments.

The ICSA model presents FF to achieve enhanced classification outcomes. It expresses a positive number to signify the improved performance of the candidate solution. The reduction of the classification rate of error is reflected as FF in this study, as expressed in Eq. (42).

$$\begin{aligned} fitness(x_i) &= ClassifierErrorRate(x_i) \\ &= \frac{number\ of\ misclassified\ samples}{Total\ number\ of\ samples} \times 100 \end{aligned} \tag{42}$$

Performance analysis

The performance analysis of the HDLMODC-ICSA methodology is examined under the Indoor object's detection dataset[34]. The dataset contains 6642 counts under 10 objects. The dataset images have a frame size of 1024 by 768 pixels. The complete details of the database are shown in Table 3.

Table 4 describes the inference time (ms), model size (m), and memory requirements (GPU) of the HDLMODC-ICSA model. The HDLMODC-ICSA model demonstrates the fastest inference time at 4.79 ms, significantly outperforming existing methods such as GoogleNet + MFF at 11.63 ms and Xception + tfidf at 11.02 ms. It also achieves a compact model size of 10 million parameters and requires only 1031 MB of GPU memory, making it highly efficient. In contrast, models like DUCA and ResNet18 + LSA have substantially larger sizes at 80 million and 77 million parameters, requiring 6383 MB and 5107 MB of GPU memory respectively. Compared to Efficient CNN + FF and G-MS2F, which demand over 7900 MB and 6900 MB of memory, the HDLMODC-

Parameter	Symbol	Description	Typical Value
Size of Populace	N	Overall crows in the populace.	30–100
MAX_ITER	T	Overall iterations for the algorithm to run.	100–500
Initial Neighborhood Radius	δ_{init}	Starting radius for search area.	0.5-1.0
Learning Rate	α_{inii}	Initial rate for learning and movement updates.	0.5-1.0
Radius Decay Factor	γ	Controls the rate of radius reduction over iterations.	0.90–0.99
Learning Decay Factor	β	Controls decay rate of learning rate.	0.90–0.99
Q-Learning Rate	η	Learning rate for Q-value updates in RL.	0.1–0.3
Discount Factor	λ	Determines future reward consideration in Q-learning.	0.8–0.99
Exploration Threshold	P_{a2}	Probability to switch between cache and recovery strategies.	0.2

Table 2. Hyperparameter settings for the ICSA method.

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 3. Details of the dataset.

Methodology	Inference Time (ms)	Model Size (m)	Memory Requirements (GPU)
HDLMODC-ICSA	4.79	10	1031
DenseNet2010-WCO-LSM	9.64	24	6669
Efficient CNN + FF	9.35	23	7975
GoogleNet + MFF	11.63	55	4336
DUCA Method	6.67	80	6383
G-MS2F Model	6.41	63	6914
ResNet18 + LSA	6.31	77	5107
Xception + tfidf	11.02	27	7343

Table 4. Comparison of methodology in terms of inference time, model size, and GPU memory requirements.

ICSA techniques reduces memory usage by over 85% while also achieving lower latency and maintaining competitive performance.

Figure 7 establishes the confusion matrix created by the HDLMODC-ICSA methodology below 80:20 and 70:30 of TRPH/TSPH. The outcomes indicate that the HDLMODC-ICSA approach effectively detects and recognizes each class.

Table 5; Fig. 8 present the indoor OD of the HDLMODC-ICSA approach under 80%TRPH and 20%TSPH. The result states that the HDLMODC-ICSA technique attained effective performance. Based on 80%TRPH, the HDLMODC-ICSA technique attained an average $accu_y$, $prec_n$, $reca_l$, $F1_{score}$, and MCC of 99.57%, 94.20%, 84.05%, 86.43%, and 87.34%, respectively. Similarly, depending on 20%TSPH, the HDLMODC-ICSA technique attained an average $accu_y$, $prec_n$, $reca_l$, $F1_{score}$, and MCC of 99.59%, 94.95%, 85.03%, 88.45%, and 88.93%, correspondingly.

Table 6; Fig. 9 exemplify the indoor OD of the HDLMODC-ICSA approach under 70%TRPH and 30%TSPH. The result states that the HDLMODC-ICSA approach has accomplished capable performance. Based on 70%TRPH, the HDLMODC-ICSA approach got an average $accu_y$, $prec_n$, $reca_l$, $F1_{score}$, and MCC of 99.41%, 91.28%, 79.83%, 83.10%, and 83.83%, respectively. Likewise, for 30%TRPH, the HDLMODC-ICSA technique got average $accu_y$, $prec_n$, $reca_l$, $F1_{score}$, and MCC of 99.51%, 91.13%, 83.08%, 85.60%, and 85.99%, correspondingly.

Figure 10 illustrates the training (TRA) $accu_y$ and validation (VAL) $accu_y$ analysis of the HDLMODC-ICSA technique under 80%TRPH and 20%TSPH. The $accu_y$ analysis is calculated across the range of 0–50 epochs. The figure highlights that the TRA and VAL $accu_y$ analysis exhibitions have an increasing tendency, which informed the capacity of the HDLMODC-ICSA approach with maximum outcomes over multiple iterations. Simultaneously, the TRA and VAL $accu_y$ remainders closer over the epochs, which indicates inferior overfitting and exhibitions maximal outcomes of the HDLMODC-ICSA approach, assuring reliable prediction on hidden samples.

Figure 11 shows the TRA loss (TRALOS) and VAL loss (VALLOS) curves of the HDLMODC-ICSA model under 80%TRPH and 20%TSPH. The loss values are computed within the range of 0–50 epochs. The TRALOS and VALLOS values establish a reducing trend, notifying the ability of the HDLMODC-ICSA method to balance a trade-off between simplification and data fitting. The constant reduction in loss values further assures a higher performance of the HDLMODC-ICSA methodology and tunes the prediction results over time.

In Fig. 12, the precision-recall (PR) graph outcomes of the HDLMODC-ICSA technique under 80%TRPH and 20%TSPH offer clarification into its outcomes by plotting PR for each class label. The figure demonstrates that the HDLMODC-ICSA technique constantly achieves maximum PR values over dissimilar classes, representing its capacity to keep up an essential section of true positive predictions between all positive predictions (precision)

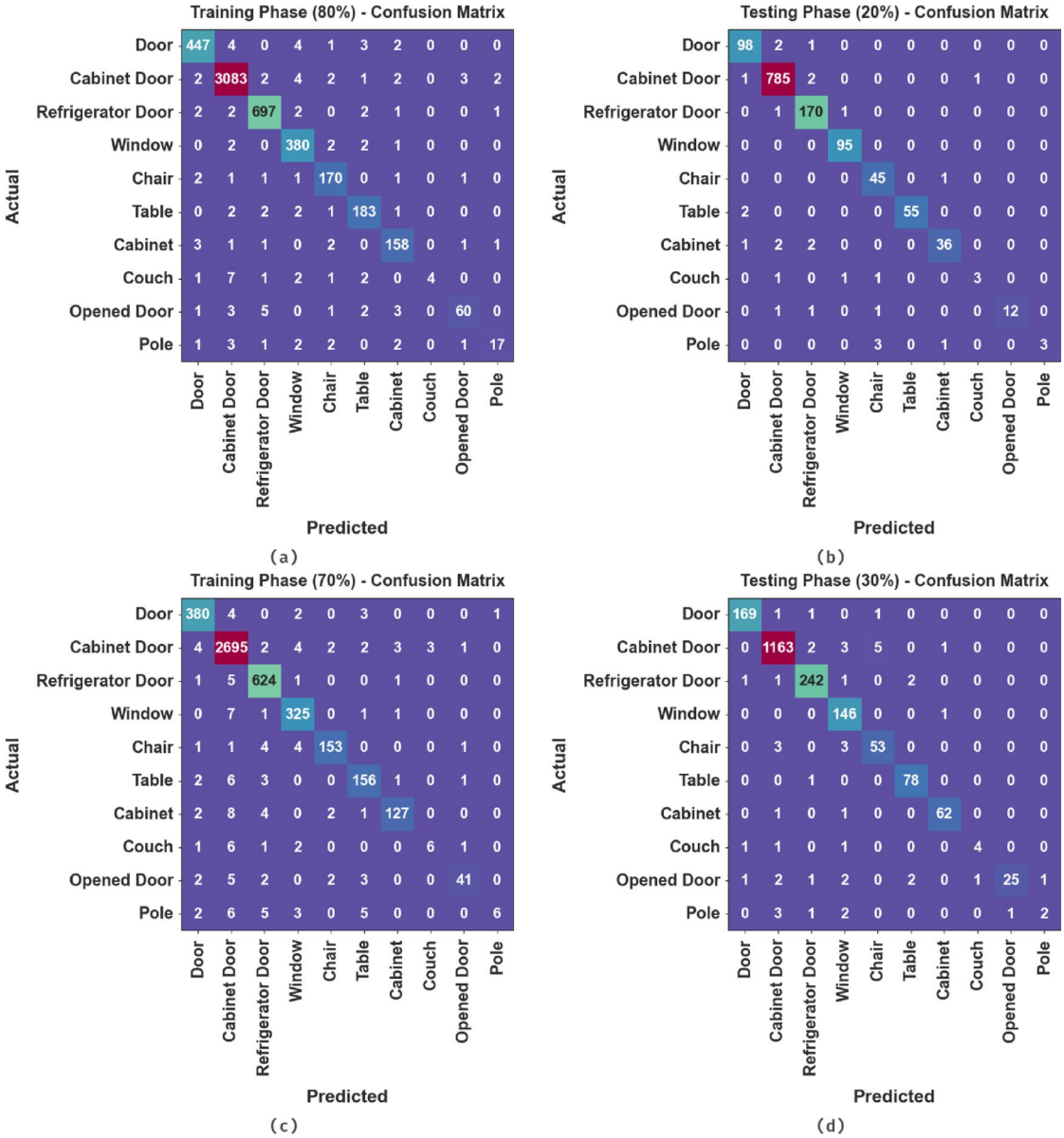

Fig. 7. Confusion matrix of (a-c) TRPH of 80% and 70% and (b-d) TSPH of 20% and 30%.

besides picking up many actual positives (recall). The established increase in PR outcomes amongst each class portrays the effectiveness of the HDLMODC-ICSA methodology in the classification procedure.

Figure 13 examines the ROC analysis of the HDLMODC-ICSA method under 80%TRPH and 20%TSPH. The outcomes suggest that the HDLMODC-ICSA technique accomplishes maximal ROC analysis across every class, indicating its ability to discriminate classes. This consistent trend of better ROC analysis across multiple classes means the capable outcomes of the HDLMODC-ICSA technique on predicting class labels highlight the robust nature of the classification process.

The comparative analysis of the HDLMODC-ICSA approach with present methodologies is demonstrated in Table 7; Fig. 14[36–38]. The simulation result shows that the HDLMODC-ICSA technique outperformed superior performances. Based on $accu_y$, the HDLMODC-ICSA approach has higher $accu_y$ of 99.59% while the DenseNet2010-WCO-LSM, Efficient CNN + FF, GoogleNet + MFF, DUCA, G-MS2F, ResNet18 + LSA, and Xception + tfidf techniques have lesser $accu_y$ of 98.78%, 95.60%, 92.92%, 94.50%, 93.22%, 89.98%, and 85.20%, respectively. Moreover, depending on $prec_n$, the HDLMODC-ICSA methodology has better $prec_n$ of 94.95 where the DenseNet2010-WCO-LSM, Efficient CNN + FF, GoogleNet + MFF, DUCA, G-MS2F, ResNet18 + LSA, and Xception + tfidf methods have lower $prec_n$ of 91.34%, 93.48%, 91.05%, 90.26%, 91.47%, 86.9%, and 84.27%, respectively.

Table 8; Fig. 15 indicates the comparison study of the HDLMODC-ICSA methodology with existing models under the Indoor OD dataset[35]. The dataset comprises 2,213 image frames annotated with seven object classes, captured under varied indoor conditions to facilitate research in OD. The comparison study demonstrates the

Class Labels	$Accu_y$	$Prec_n$	$Reca_l$	$F1_{score}$	MCC	Kappa
TRPH (80%)						
Door	99.51	97.39	96.96	97.17	96.91	97.54
Cabinet Door	99.19	99.20	99.42	99.31	98.33	99.04
Refrigerator Door	99.57	98.17	98.59	98.38	98.13	98.79
Window	99.55	95.72	98.19	96.94	96.70	97.40
Chair	99.64	93.41	96.05	94.71	94.53	95.28
Table	99.62	93.85	95.81	94.82	94.63	95.39
Cabinet	99.59	92.40	94.61	93.49	93.28	94.06
Couch	99.74	100.00	22.22	36.36	47.08	47.60
Opened Door	99.60	90.91	80.00	85.11	85.09	85.60
Pole	99.70	80.95	58.62	68.00	68.75	69.55
Average	**99.57**	**94.20**	**84.05**	**86.43**	**87.34**	**97.40**
TSPH (20%)						
Door	99.47	96.08	97.03	96.55	96.27	97.03
Cabinet Door	99.17	99.12	99.49	99.30	98.28	98.98
Refrigerator Door	99.40	96.59	98.84	97.70	97.36	98.13
Window	99.85	97.94	100.00	98.96	98.88	99.60
Chair	99.55	90.00	97.83	93.75	93.60	94.16
Table	99.85	100.00	96.49	98.21	98.15	98.78
Cabinet	99.47	94.74	87.80	91.14	90.94	91.53
Couch	99.70	75.00	50.00	60.00	61.10	61.62
Opened Door	99.77	100.00	80.00	88.89	89.34	90.04
Pole	99.70	100.00	42.86	60.00	65.37	65.89
Average	**99.59**	**94.95**	**85.03**	**88.45**	**88.93**	**89.58**

Table 5. Indoor OD of HDLMODC-ICSA model under 80%TRPH and 20%TSPH.

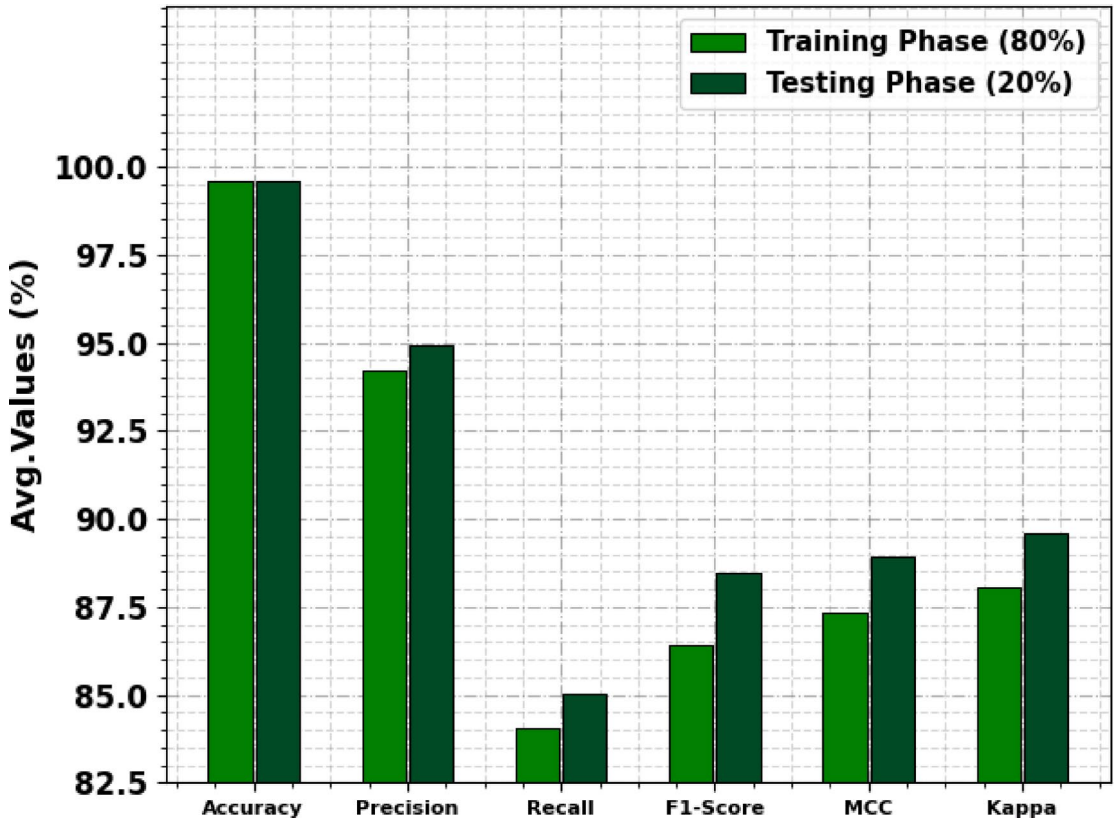

Fig. 8. Average of HDLMODC-ICSA model under 80%TRPH and 20%TSPH.

Class Labels	$Accu_y$	$Prec_n$	$Reca_l$	$F1_{score}$	MCC	Kappa
TRPH (70%)						
Door	99.46	96.20	97.44	96.82	96.52	97.16
Cabinet Door	98.52	98.25	99.23	98.74	96.95	97.52
Refrigerator Door	99.35	96.59	98.73	97.65	97.29	97.90
Window	99.44	95.31	97.01	96.15	95.86	96.49
Chair	99.63	96.23	93.29	94.74	94.56	95.14
Table	99.40	91.23	92.31	91.76	91.45	92.03
Cabinet	99.51	95.49	88.19	91.70	91.52	92.30
Couch	99.70	66.67	35.29	46.15	48.38	49.08
Opened Door	99.61	91.11	74.55	82.00	82.23	82.82
Pole	99.53	85.71	22.22	35.29	43.51	44.31
Average	**99.41**	**91.28**	**79.83**	**83.10**	**83.83**	**84.48**
TSPH (30%)						
Door	99.70	98.26	98.26	98.26	98.09	98.72
Cabinet Door	98.85	98.98	99.06	99.02	97.62	98.21
Refrigerator Door	99.45	97.58	97.98	97.78	97.46	98.04
Window	99.30	91.82	99.32	95.42	95.13	95.87
Chair	99.40	89.83	89.83	89.83	89.52	90.07
Table	99.75	95.12	98.73	96.89	96.78	97.56
Cabinet	99.80	96.88	96.88	96.88	96.77	97.54
Couch	99.80	80.00	57.14	66.67	67.52	68.30
Opened Door	99.45	96.15	71.43	81.97	82.63	83.14
Pole	99.60	66.67	22.22	33.33	38.34	39.05
Average	**99.51**	**91.13**	**83.08**	**85.60**	**85.99**	**86.65**

Table 6. Indoor OD of HDLMODC-ICSA model under 70%TRPH and 30%TSPH.

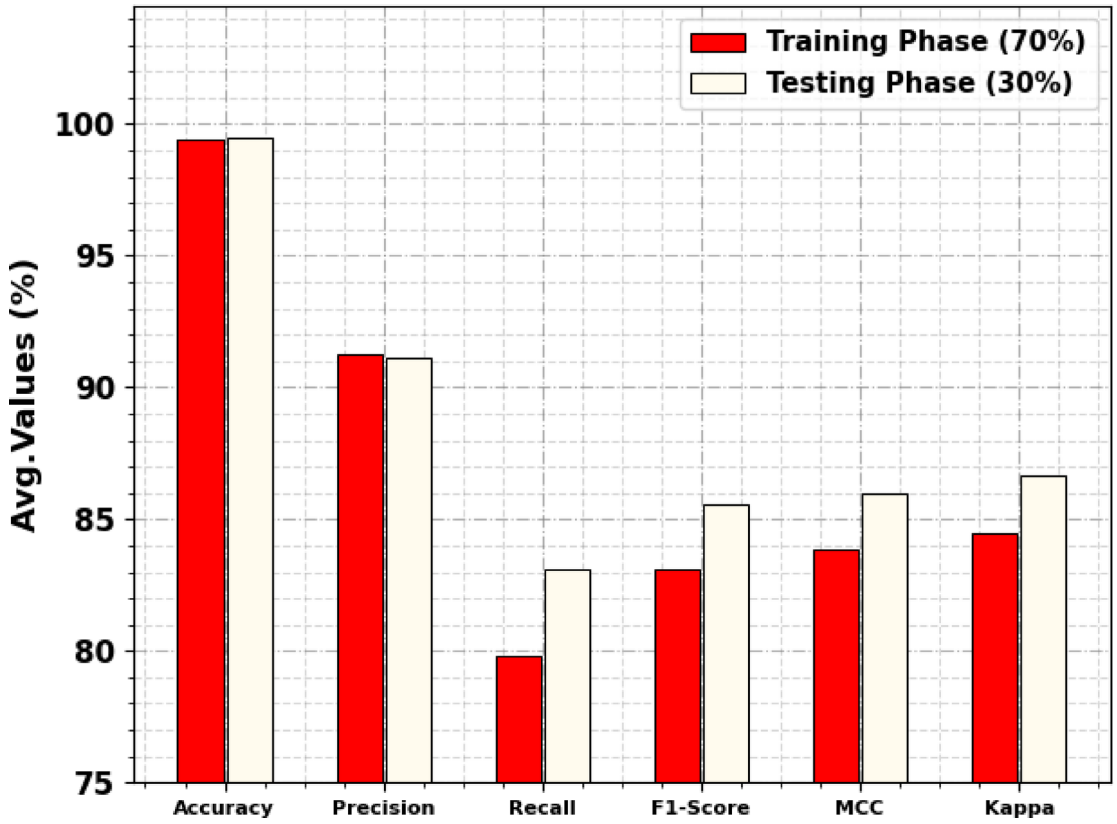

Fig. 9. Average of HDLMODC-ICSA model under 70%TRPH and 30%TSPH.

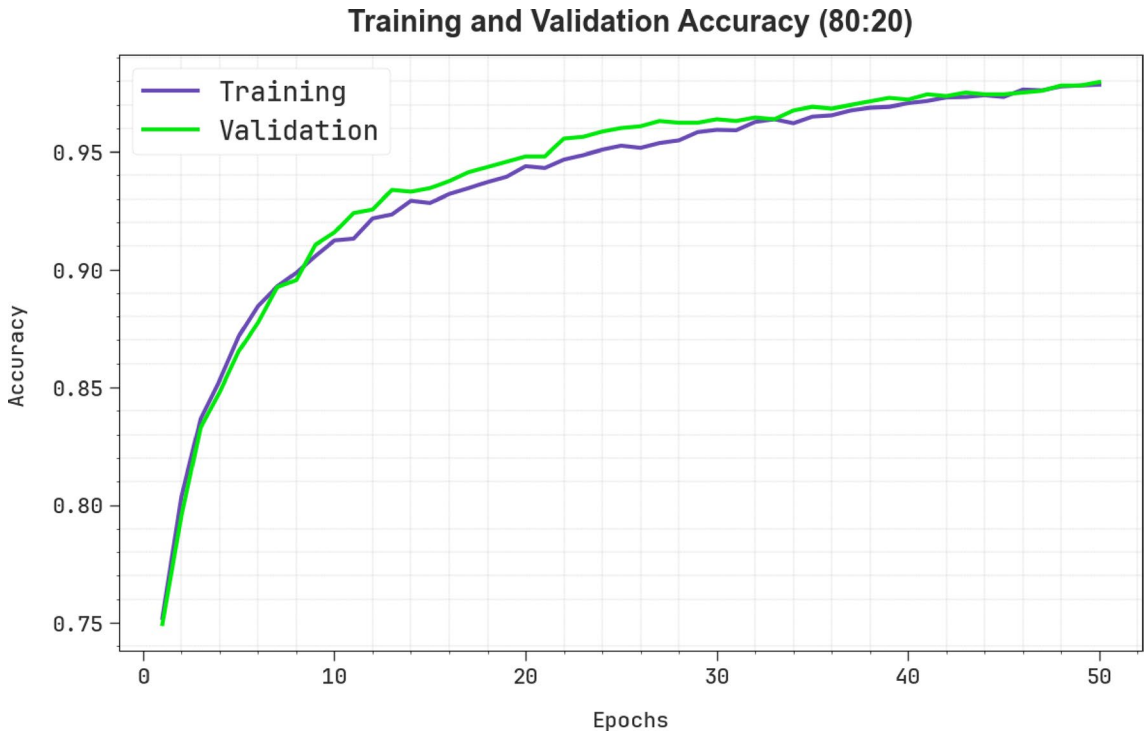

Fig. 10. $Accu_y$ analysis of HDLMODC-ICSA model under 80%TRPH and 20%TSPH.

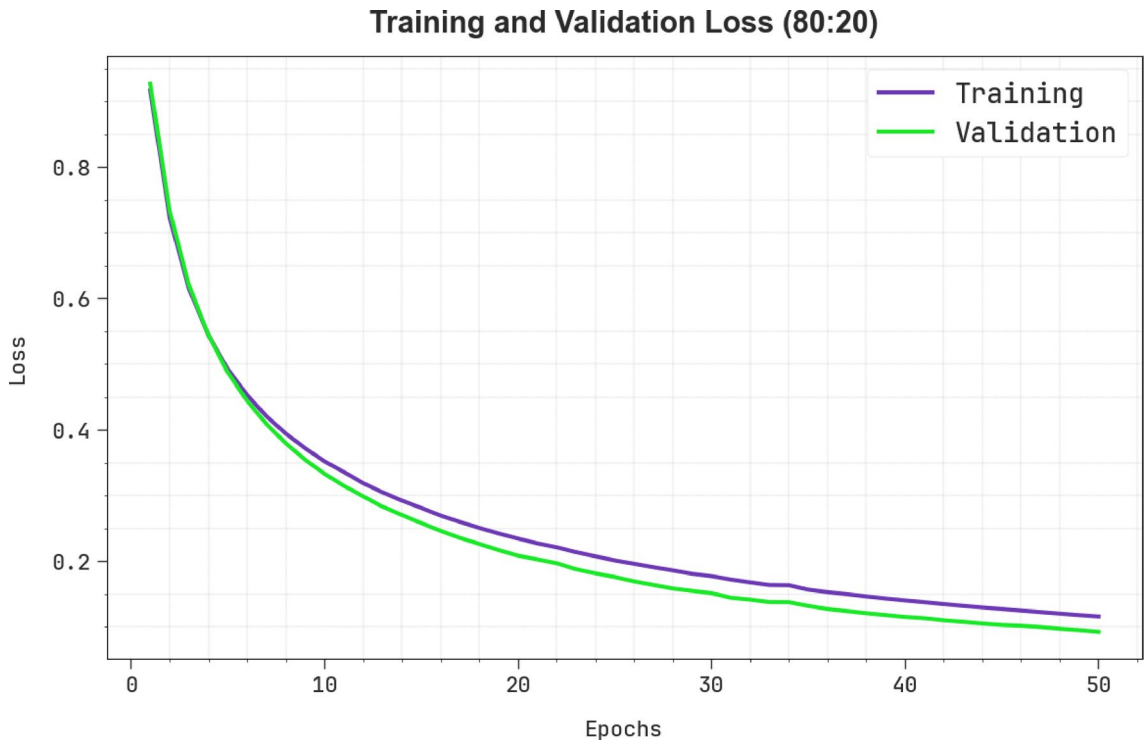

Fig. 11. Loss graph of HDLMODC-ICSA model under 80%TRPH and 20%TSPH.

effectiveness of the HDLMODC-ICSA methodology over various established OD models[39,40]. The HDLMODC-ICSA methodology achieves the highest $accu_y$ at 99.59% and leads in both $prec_n$ and $reca_l$ with 94.95% and 85.03% respectively, resulting in an $F1_{score}$ of 88.45%. While DC-SPP-YOLO and YOLOv5m also perform well with $accu_y$ of 99.25% and 98.71%, their $prec_n$ and $reca_l$ values fall short compared to HDLMODC-ICSA

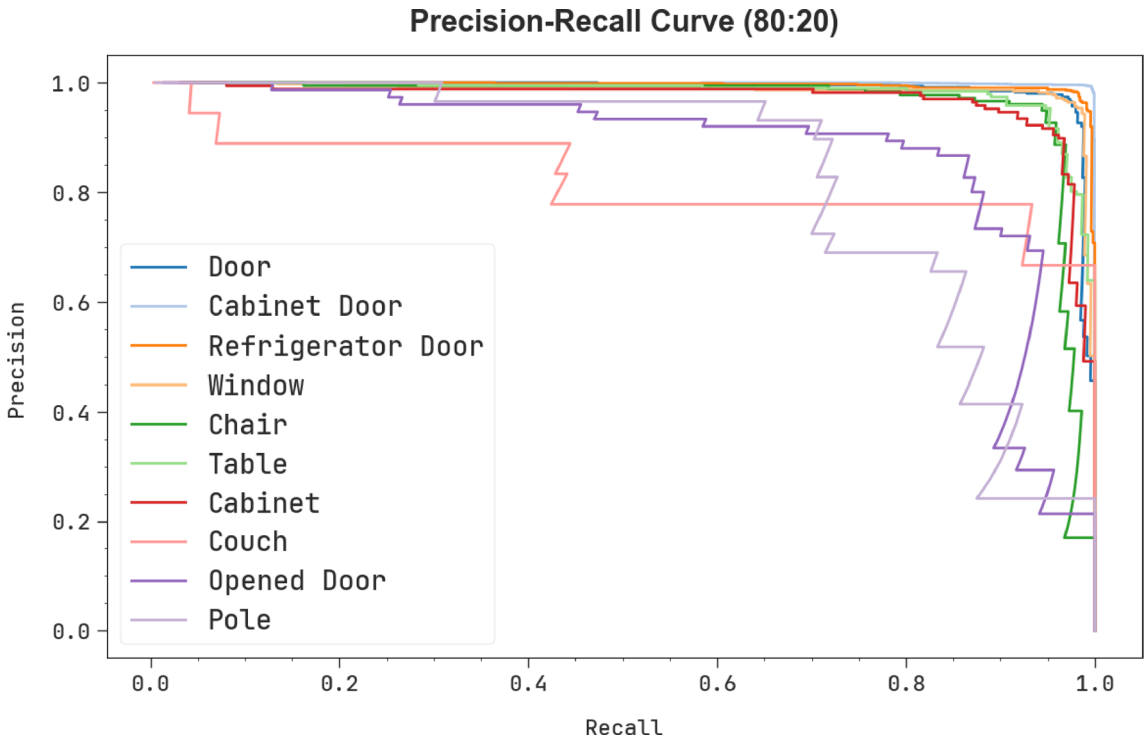

Fig. 12. PR analysis of HDLMODC-ICSA model under 80%TRPH and 20%TSPH.

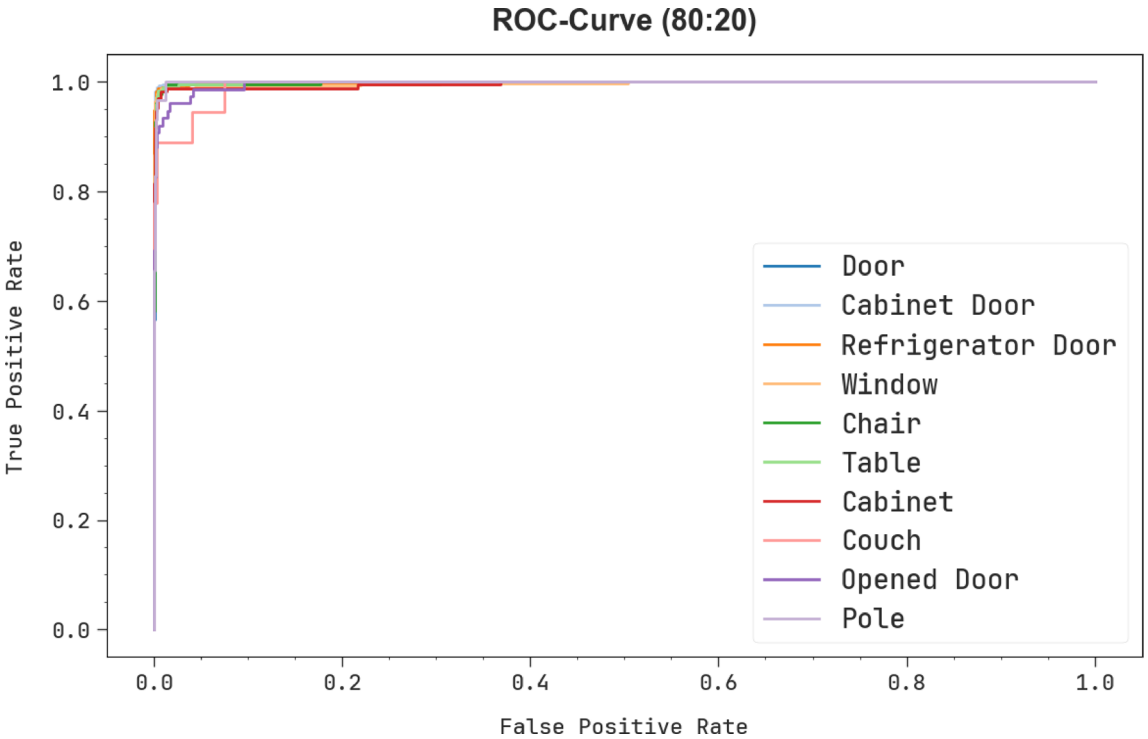

Fig. 13. ROC graph of HDLMODC-ICSA model under 80%TRPH and 20%TSPH.

Methodology	$Accu_y$	$Prec_n$	$Reca_l$	$F1_{score}$
HDLMODC-ICSA	99.59	94.95	85.03	88.45
DenseNet2010-WCO-LSM	98.78	91.34	80.15	77.81
Efficient CNN + FF	95.60	93.48	79.95	73.97
GoogleNet + MFF	92.92	91.05	82.26	81.84
DUCA Method	94.50	90.26	82.97	81.29
G-MS2F Model	93.22	91.47	78.57	76.83
ResNet18 + LSA	89.98	86.95	73.41	77.83
Xception + tfidf	85.20	84.27	74.52	75.28

Table 7. Comparative results of HDLMODC-ICSA model with existing techniques [36 – 28].

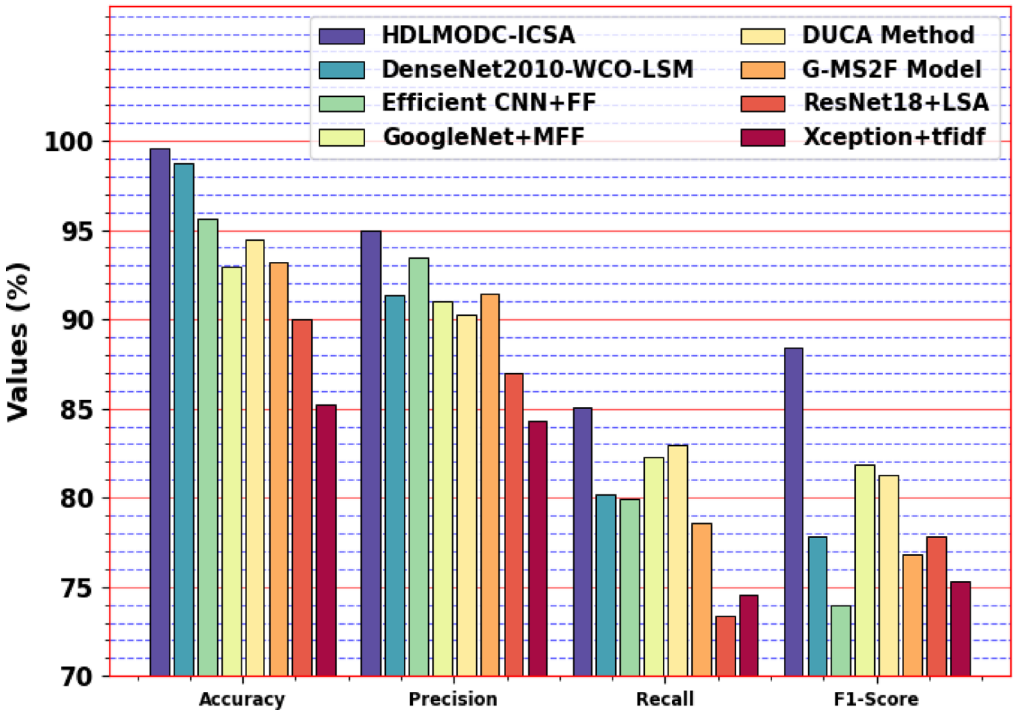

Fig. 14. Comparative analysis of HDLMODC-ICSA model with existing techniques.

Technique	$Accu_y$	$Prec_n$	$Reca_l$	$F1_{score}$
YOLOv5n	95.94	93.46	84.98	87.11
YOLOv5m	98.71	91.79	80.56	88.15
JET-Net	92.48	91.50	82.76	83.46
Tiny-YOLO	93.21	93.14	81.44	84.10
Mask R-CNN	93.47	92.69	81.08	84.64
DC-SPP-YOLO	99.25	90.38	80.28	83.51
AT-LI-YOLO	93.77	90.54	84.42	85.87
HDLMODC-ICSA	99.59	94.95	85.03	88.45

Table 8. Comparison analysis of HDLMODC-ICSA methodology with existing models.

technique. For instance, YOLOv5m records a lower $reca_l$ of 80.56%, and DC-SPP-YOLO exhibits a $prec_n$ of 90.38% and $F1_{score}$ of 83.51%. Other models like AT-LI-YOLO, Mask R-CNN, and Tiny-YOLO achieve good balance but still lag in overall performance. These results confirm that the HDLMODC-ICSA technique presents a more robust and accurate solution for OD tasks.

In Table 9; Fig. 16, the comparative outcomes of the HDLMODC-ICSA approach are identified in Mean IoU. The outcomes imply that the HDLMODC-ICSA methodology gets superior outcomes. Depend on mean IoU, the

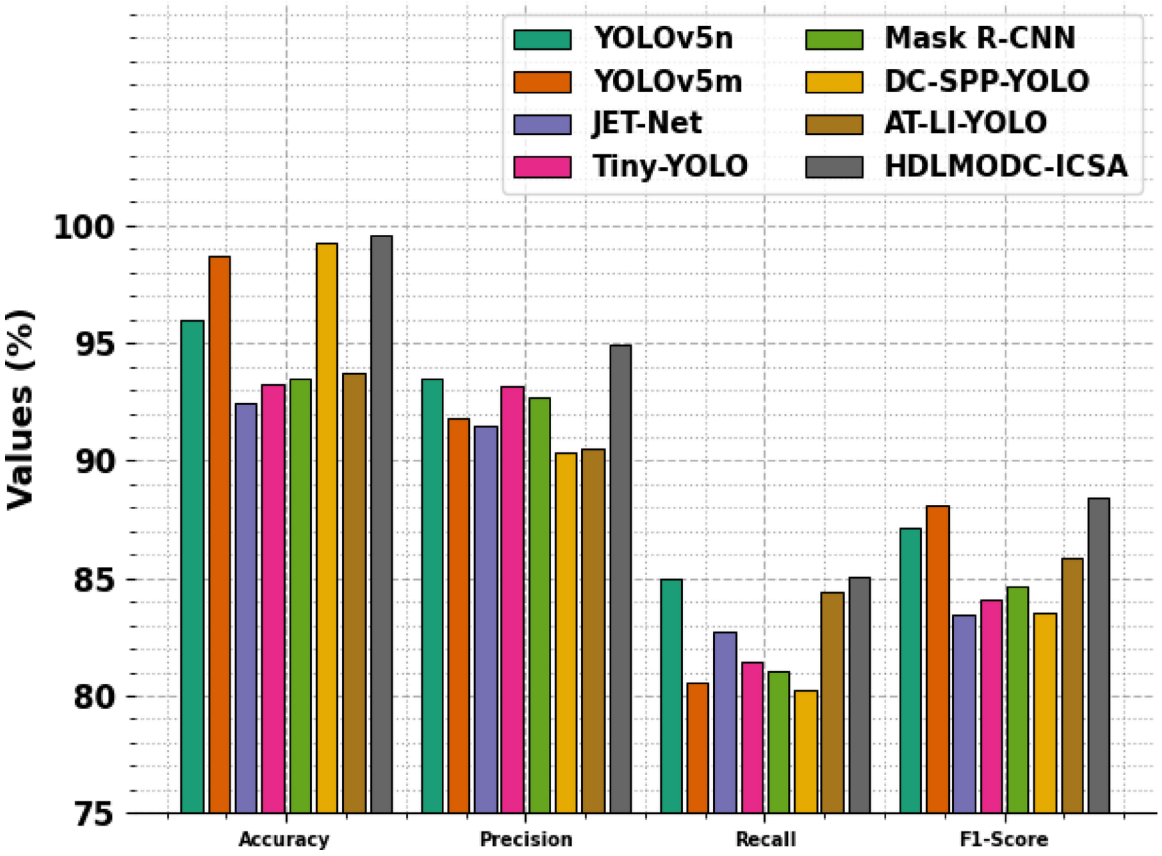

Fig. 15. Comparison analysis of HDLMODC-ICSA methodology with existing models.

Methodology	Mean IoU (%)
HDLMODC-ICSA	79.48
DenseNet2010-WCO-LSM	72.88
Efficient CNN + FF	72.33
GoogleNet + MFF	68.11
DUCA Method	73.48
G-MS2F Model	63.11
ResNet18 + LSA	72.17
Xception + tfidf	69.33

Table 9. Mean IoU outcome of HDLMODC-ICSA approach with existing methods.

HDLMODC-ICSA methodology provides higher value of 79.48%, whereas DenseNet2010-WCO-LSM, Efficient CNN + FF, GoogleNet + MFF, DUCA, G-MS2F, ResNet18 + LSA, and Xception + tfidf models have gained inferior mean IoU values of 72.88%, 72.33%, 68.11%, 73.48%, 63.11%, 72.17%, and 69.33%, correspondingly.

Table 10; Fig. 17 inspect the running time (RT) result of the HDLMODC-ICSA technique with existing approaches. The table values specify that the HDLMODC-ICSA approach has attained a minimal RT of 2.10 s. Whereas the existing methodologies such as DenseNet2010-WCO-LSM, Efficient CNN + FF, GoogleNet + MFF, DUCA, G-MS2F, ResNet18 + LSA, and Xception + tfidf approaches have achieved higher RT values of 3.35 s, 3.13 s, 5.15 s, 4.35 s, 4.06 s, 5.50 s, and 5.26 s, correspondingly.

Table 11; Fig. 18 specifies the ablation study of the HDLMODC-ICSA technique with existing models. The HDLMODC-ICSA technique achieves the highest $accu_y$ at 99.59%, $prec_n$ at 94.95%, $reca_l$ at 85.03%, and $F1_{score}$ at 88.45%, outperforming existing models. For instance, ABS-BiLSTM records slightly lower values with an $accu_y$ of 99.04% and 87.95% $F1_{score}$, while ICSA and LeNet-5 follow with $accu_y$ of 98.48% and 97.74% respectively. The Faster R-CNN model exhibits the lowest performance, attaining an $accu_y$ of 97.07% and an $F1_{score}$ of 86.14%. These results confirm that the HDLMODC-ICSA method not only delivers high classification accuracy but also maintains a robust balance between $prec_n$ and $reca_l$, making it more reliable for real-world applications.

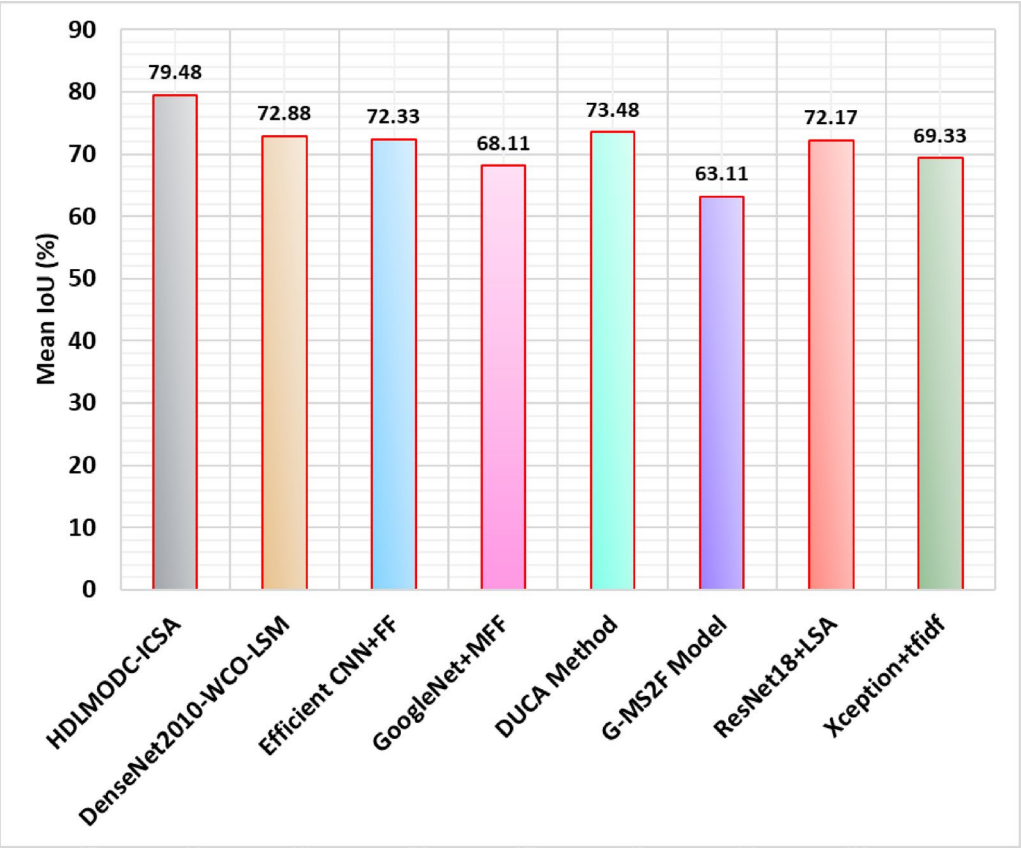

Fig. 16. Mean IoU outcome of HDLMODC-ICSA approach with existing methods.

Methods	RT (sec)
HDLMODC-ICSA	2.10
DenseNet2010-WCO-LSM	3.35
Efficient CNN + FF	3.13
GoogleNet + MFF	5.15
DUCA Method	4.35
G-MS2F Model	4.06
ResNet18 + LSA	5.50
Xception + tfidf	5.26

Table 10. RT outcome of HDLMODC-ICSA model with existing methods.

Table 12; Fig. 19 indicates the error analysis of the HDLMODC-ICSA methodology with existing models. The error analysis indicates that while the HDLMODC-ICSA methodology illustrates relatively low performance metrics across $accu_y$, $prec_n$, $reca_l$, and $F1_{score}$, other models such as Xception + tfidf and ResNet18 + LSA outperform it significantly. The lower performance of the HDLMODC-ICSA model may be attributed to challenges in feature extraction or the inability of the model to effectively capture the complex relationships within the data. On the contrary, models like DenseNet2010-WCO-LSM and Efficient CNN + FF exhibit improved recall, but they still fall short of attaining a balanced performance across all metrics. This suggests that while these models capture some useful features, they may still struggle with precision and F1 score, representing room for improvement in error reduction and handling data imbalances.

Table 13 depicts the performance of the HDLMODC-ICSA model under varying training and testing splits, highlighting consistent and robust results across both metrics. With 80% training data, the model achieves a weighted-F1 score of 97.48% and a macro-F1 score of 77.67%, while 70% training presents slightly lower weighted-F1 at 96.77% but a higher macro-F1 of 79.54%. On the testing side, the model outperforms, recording 97.77% weighted-F1 and 82.45% macro-F1 with 20% test data, and additionally improving to 97.45% weighted-F1 and 88.27% macro-F1 when tested on 30% data. These results reflect the robustness and generalization capability of the model across diverse data splits.

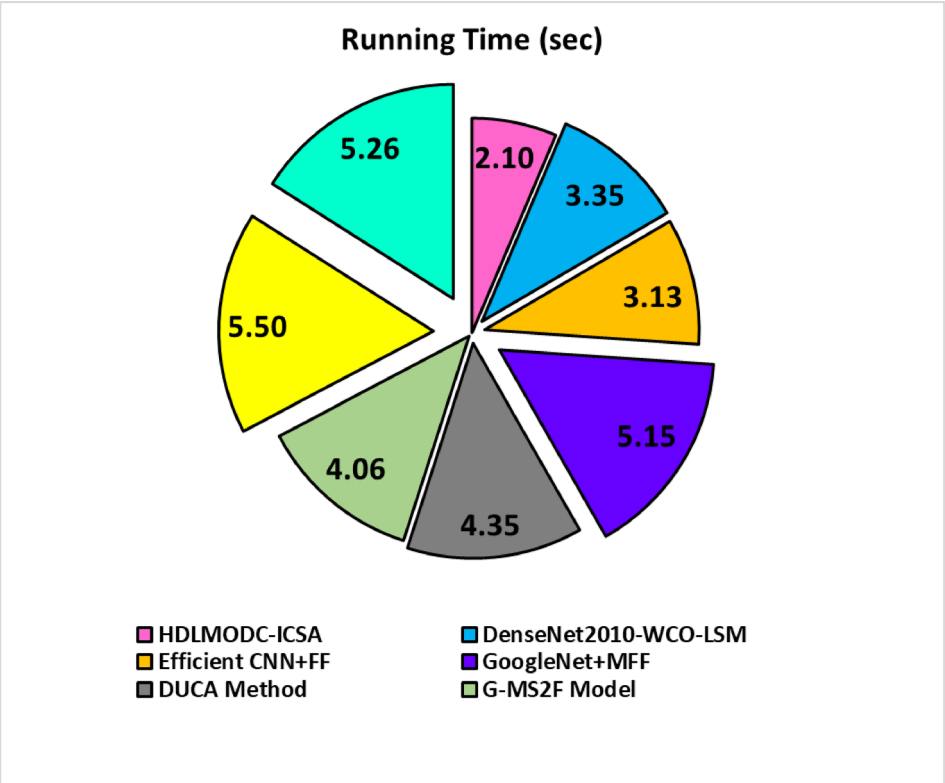

Fig. 17. RT outcome of HDLMODC-ICSA approach with existing methods.

Methodology	$Accu_y$	$Prec_n$	$Reca_l$	$F1_{score}$
HDLMODC-ICSA	99.59	94.95	85.03	88.45
ABS-BiLSTM	99.04	94.24	84.50	87.95
ICSA	98.48	93.59	83.92	87.24
LeNet-5	97.74	93.04	83.18	86.70
Faster R-CNN	97.07	92.43	82.50	86.14

Table 11. Result analysis of the ablation study of HDLMODC-ICSA technique with existing methods.

Conclusion

This study develops and designs an HDLMODC-ICSA method. The HDLMODC-ICSA method primarily focuses on an accurate and real-time object recognition method to aid visually challenged persons. In the initial stage, the image pre-processing stage applies MF to remove noise or distortions to make the image more transparent. Additionally, the OD process is executed by the Faster R-CNN model to generate precise region proposals and detect objects within images efficiently. Furthermore, the proposed HDLMODC-ICSA method employs the Improved LeNet-5 model to extract meaningful and discriminative features from the identified regions. For OD and classification, the hybrid of the ABS-BiLSTM technique is used. Finally, the hyperparameter selection of the ABS-BiLSTM model is performed by implementing ICSA. The efficiency of the HDLMODC-ICSA approach is validated by comprehensive studies using the Indoor objects detection dataset. The comparison study of the HDLMODC-ICSA approach demonstrated a superior accuracy value of 99.59% over existing techniques. The limitations of the HDLMODC-ICSA approach comprise various factors that could be improved in future work. Initially, the performance of the model may degrade when dealing with highly noisy or highly intrinsic datasets, as it depends heavily on pre-processing and feature extraction steps. Furthermore, the computational cost could be prohibitive for real-time applications, specifically when dealing with large-scale datasets. Another limitation is the dependability of the model on specific architectures, which might not generalize well to all domains or types of input data. Future work may concentrate on optimizing the system for faster inference times, exploring more generalized feature extraction techniques, or integrating unsupervised learning methods to mitigate the dependency on labeled data. Furthermore, improving the robustness of the technique to varying conditions and expanding its capability to handle multimodal data could be promising directions for improvement.

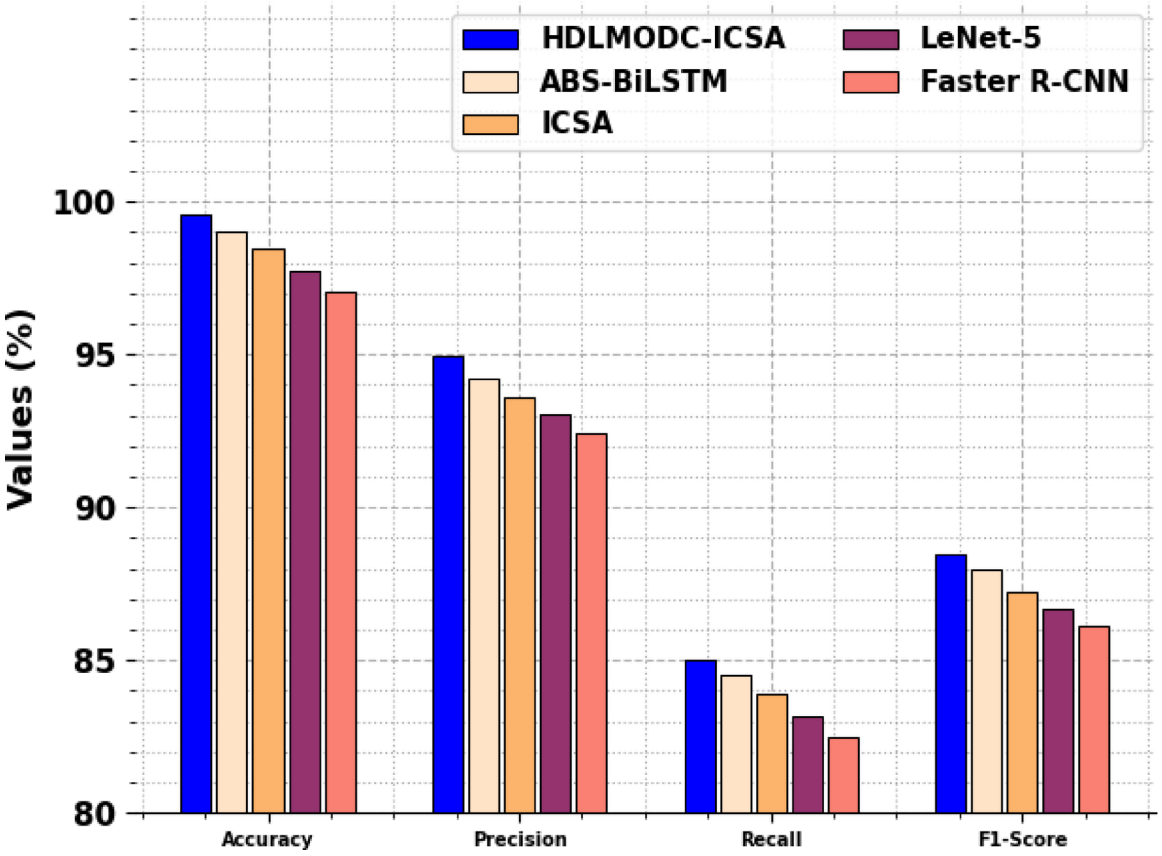

Fig. 18. Result analysis of the ablation study of HDLMODC-ICSA technique with existing methods.

Methodology	$Accu_y$	$Prec_n$	$Reca_l$	$F1_{score}$
HDLMODC-ICSA	0.41	5.05	14.97	11.55
DenseNet2010-WCO-LSM	1.22	8.66	19.85	22.19
Efficient CNN + FF	4.40	6.52	20.05	26.03
GoogleNet + MFF	7.08	8.95	17.74	18.16
DUCA Method	5.50	9.74	17.03	18.71
G-MS2F Model	6.78	8.53	21.43	23.17
ResNet18 + LSA	10.02	13.05	26.59	22.17
Xception + tfidf	14.80	15.73	25.48	24.72

Table 12. Error analysis of HDLMODC-ICSA methodology with existing models.

Splits	Weighted-F1	Macro-F1
Training (80%)	0.9748	0.7767
Training (70%)	0.9677	0.7954
Testing (20%)	0.9777	0.8245
Testing (30%)	0.9745	0.8827

Table 13. Performance evaluation based on different training and testing data splits using weighted-f1 and macro-f1 scores.

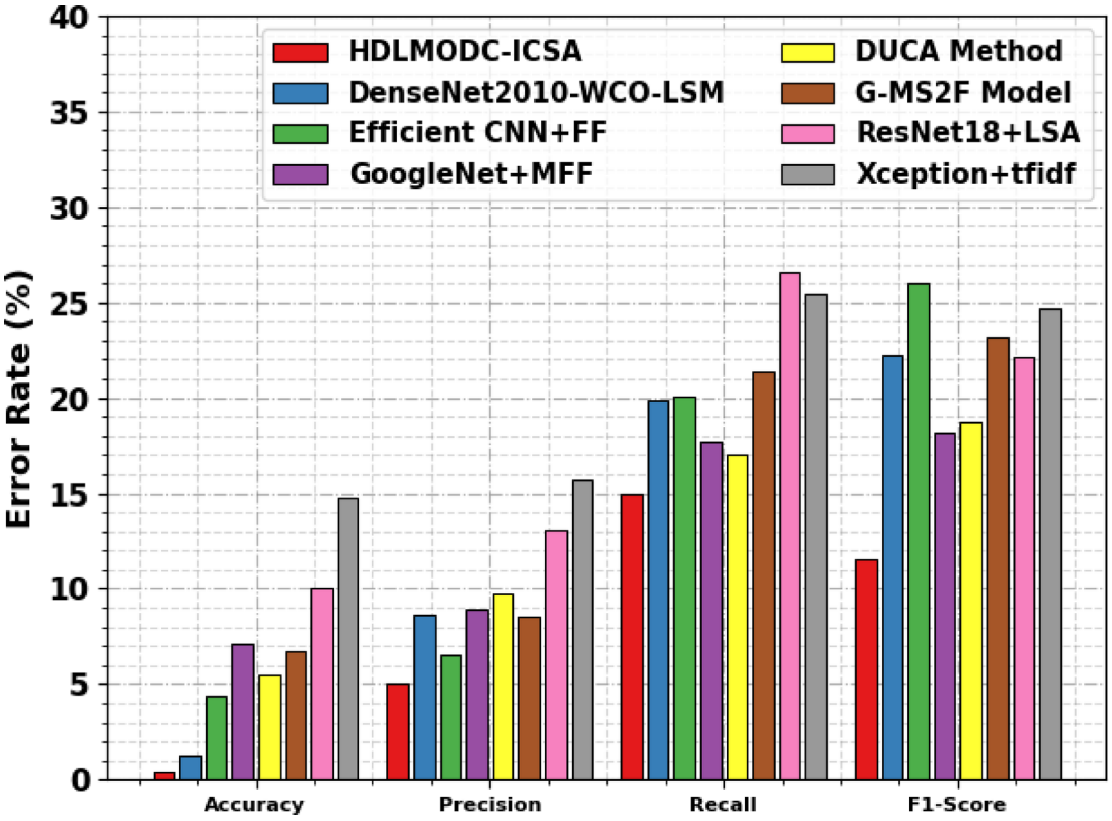

Fig. 19. Error analysis of HDLMODC-ICSA methodology with existing models.

Data availability

The data supporting this study's findings are openly available in the Kaggle repository at https://www.kaggle.com/datasets/thepbordin/indoor-object-detection, https://zenodo.org/records/2654485#.Y9FTiRXMJD9, reference number [34-35].

Received: 26 February 2025; Accepted: 11 August 2025
Published online: 14 August 2025

References

1. 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).
2. Trabelsi, R. et al. Indoor object recognition in RGBD images with complex-valued neural networks for visually-impaired people. *Neurocomputing* **330**, 94–103 (2019).
3. Islam, R. B., Akhter, S., Iqbal, F., Rahman, M. S. U. & Khan, R. Deep learning based object detection and surrounding environment description for visually impaired people. *Heliyon*, **9**(6). (2023).
4. Kaur, B. & Bhattacharya, J. Scene perception system for visually impaired based on object detection and classification using multimodal deep convolutional neural network. *J. Electron. Imaging*. **28** (1), 013031–013031 (2019).
5. Rahman, M. A. & Sadi, M. S. IoT enabled automated object recognition for the visually impaired. *Comput. Methods Programs Biomed. Update*. **1**, 100015 (2021).
6. 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).
7. Prashar, D., Chakraborty, G. S. & Jha, S. Energy efficient laser based embedded system for blind turn traffic control. *J. Cybersecur. Inform. Manage.* 35–43. https://doi.org/10.54216/jcim.020201 (2020).
8. Joshi, R. C., Yadav, S., Dutta, M. K. & Travieso-Gonzalez, C. M. Efficient multi-object detection and smart navigation using artificial intelligence for visually impaired people. *Entropy*, *22*(9), p.941. (2020).
9. Cordeiro, N. H. & Pedrino, E. C. A new methodology applied to dynamic object detection and tracking systems for visually impaired people. *Comput. Electr. Eng.* **77**, 61–71 (2019).
10. Salama, R. Enhancing object detection and classification using white shark optimization with deep learning on remote sensing images. *Fusion: Pract. Appl.* **2**, 147–147 (2025).
11. Abidi, M. H., Alkhalefah, H. & Siddiquee, A. N. Enhancing Navigation and Object Recognition for Visually Impaired Individuals: A Gradient Support Vector Boosting-based Crossover Golden Jackal Algorithm Approach. *Journal of Disability Research*, *3*(5), p.20240057. (2024).
12. Baskar, A., Kumar, T. G. & Samiappan, S. A vision system to assist visually challenged people for face recognition using multi-task cascaded convolutional neural network (MTCNN) and local binary pattern (LBP). *J. Ambient Intell. Humaniz. Comput.* **14** (4), 4329–4341 (2023).
13. Masal, K. M., Bhatlawande, S. & Shingade, S. D. An integrated region proposal and Spatial information guided Convolution network-based object recognition for visually impaired persons' indoor assistive navigation. *Imaging Sci. J.* **72** (7), 884–897 (2024).

14. Priyanka, R. et al. November. Robust Object Detection and Tracking Model for Visually Impaired People Using Deep Convolution Neural Network Model. In *2023 International Conference on Sustainable Communication Networks and Application (ICSCNA)* (pp. 1593–1598). IEEE. (2023).
15. Bai, Z. et al. *Enhanced Lightweight Infrared Object Detection Algorithm for Assistive Navigation in Visually Impaired Individuals* (IET Image Processing, 2024).
16. Gupta, C., Gill, N. S., Gulia, P. & Chatterjee, J. M. A novel fine-tuned YOLOv6 transfer learning model for real-time object detection. *Journal of Real-Time Image Processing*, *20*(3), p.42. (2023).
17. Ikram, S., Bajwa, I., Gyawali, S., Ikram, A. & Alsubaie, N. A IoT-enabled Obstacle Detection and Recognition Technique for Blind Persons. (2024).
18. Tiwary, T. & Mahapatra, R. P. Enhancement in web accessibility for visually impaired people using hybrid deep belief network–bald eagle search. *Multimedia Tools Appl.* **82** (16), 24347–24368 (2023).
19. Sindhu, B., Preethi, B. B., Lakshmi, S. L. V., Reddy, B. P. & Kiran, S. S. Voice-Assisted Artificial Intelligence Based Question Answering System for the Visually Impaired. *Algorithms in Advanced Artificial Intelligence*, p.135. (2025).
20. Dang, B. et al. May. Real-time pill identification for the visually impaired using deep learning. In *2024 6th International Conference on Communications, Information System and Computer Engineering (CISCE)* (pp. 552–555). IEEE. (2024).
21. Ayadi, M., Masmoudi, N., Almuqren, L., Aljohani, R. O. & Alshahrani, H. S. Empowering Accessibility in Handwritten Arabic Text Recognition for Visually Impaired Individuals through Optimized Generative Adversarial Network (GAN) model. *Journal of Disability Research*, *4*(1), p.20240110. (2025).
22. Alruwaili, M., Siddiqi, M. H., Atta, M. N. & Arif, M. Deep learning and ubiquitous systems for disabled people detection using YOLO models. *Computers in Human Behavior*, *154*, p.108150. (2024).
23. Kotis, K., Angoura, E. & Lyngri, E. I. Emerging technologies in smart libraries for visually impaired people: challenges and design considerations. *ACM J. Comput. Cult. Heritage* (2025).
24. Saini, M. & Sengupta, E. Artificial intelligence inspired fog-cloud-based visual-assistance framework for blind and visually-impaired people. *Multimedia Tools Applications*, pp.1–32. (2024).
25. Sumithra, S., Ponnrajakumari, M. & Dharshini, T. Breaking barriers: quantum Computing-enhanced smart ATM for Multi-Sensory impairment people using OpenCV. In Multidisciplinary Applications of AI and Quantum Networking (343–356). IGI Global. (2025).
26. Alshehri, M. M., Sharma, S. K., Gupta, P. & Shah, S. R. Empowering the visually impaired: Translating handwritten digits into spoken language with HRNN-GOA and Haralick features. *Journal of Disability Research*, *3*(1), p.20230051. (2024).
27. Qi, X., Li, H., Li, L. & Wu, Z. EmoAssist: Emotional Assistant for Visual Impairment Community. *arXiv preprint arXiv:2502.09285*. (2025).
28. Pongiannan, R. K., Franklin, J. & Richard Pravin, A. December. Object Detection Using Deep Learning for Blind People with Voice Feedback. In *2024 International Conference on System, Computation, Automation and Networking (ICSCAN)* (pp. 1–5). IEEE. (2024).
29. Kalita, D. & Lyakhov, P. Moving object detection based on a combination of Kalman filter and median filtering. *Big Data and Cognitive Computing*, *6*(4), p.142. (2022).
30. Hong, Q., Dong, H., Deng, W. & Ping, Y. Education robot object detection with a brain-inspired approach integrating Faster R-CNN, YOLOv3, and semi-supervised learning. *Frontiers in Neurorobotics*, *17*, p.1338104. (2024).
31. Wei, R. & Lin, Q. Intelligent Evaluation of Tourism Competitiveness Based on Improved Lenet-5 Network Model. (2024).
32. Rautaray, J., Panigrahi, S. & Nayak, A. K. Integrating particle swarm optimization with Backtracking search optimization feature extraction with two-dimensional convolutional neural network and attention-based stacked bidirectional long short-term memory classifier for effective single and multi-document summarization. *PeerJ Comput. Sci.* **10**, e2435 (2024).
33. Liu, L., Dai, L., Mao, X., Chen, Y. & Jing, Y. Research on Gas Emission Prediction Based on KPCA-ICSA-SVR. *Processes*, *12*(12), p.2655. (2024).
34. https://www.kaggle.com/datasets/thepbordin/indoor-object-detection
35. https://zenodo.org/records/2654485#.Y9FTiRXMJD9
36. Surendran, R., Chihi, I., Anitha, J. & Hemanth, D. J. Indoor Scene Recognition: An Attention-Based Approach Using Feature Selection-Based Transfer Learning and Deep Liquid State Machine. *Algorithms*, *16*(9), p.430. (2023).
37. Sitaula, C., Xiang, Y., Zhang, Y., Lu, X. & Aryal, S. Indoor image representation by high-level semantic features. *IEEE Access.* **7**, 84967–84979 (2019).
38. Heikel, E. & Espinosa-Leal, L. Indoor scene recognition via object detection and TF-IDF. *Journal of Imaging*, *8*(8), p.209. (2022).
39. Azurmendi, I., Zulueta, E., Lopez-Guede, J. M. & González, M. Simultaneous object detection and distance estimation for indoor autonomous vehicles. *Electronics*, *12*(23), p.4719. (2023).
40. Ye, Y. et al. Dynamic and real-time object detection based on deep learning for home service robots. *Sensors*, *23*(23), p.9482. (2023).

Acknowledgements

The authors extend their appreciation to the King Salman center for Disability Research for funding this work through Research Group no KSRG-2024-070.

Author contributions

Alaa O. Khadidos: Conceptualization, methodology, validation, investigation, writing—original draft preparation, fundingAyman Yafoz: Conceptualization, methodology, writing—original draft preparation, writing—review and editing.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

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

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.