ACTA SCIENTIAE

ISSN: 2178-7727

Dynamic Object Prioritization and Contextual Scene Analysis in Assistive Navigation using a Hybrid Self-Attentive Convolutional Bi-GRU-Net Ar-chitecture

Komal Mahadeo Masal[1], Shripad Bhatlawande[2], Sachin Dattatraya Shingade[3]

[1] DOT SPPU, PICT Pune, Maharashtra India,
kmmasal@pict.edu
[2] Vishwakarma Institute of Technology VIT Pune, Maharashtra, India,
shripad.bhatlawande@vit.edu
[3]Pune Institute of Computer Technology PICT Pune, Maharashtra, IIT Patna India,
sachin_pa2503mth305@iitp.ac.in

ABSTRACT

The provision of autonomous mobility for the visually impaired necessitates the integration of high-fidelity en-vironmental perception and intelligent hazard assessment. This paper presents a novel deep learning paradigm, 'Dynamic object prioritization and contextual scene analysis in assistive navigation using a hybrid self attentive convolutional Bi-GRU-Net architecture', engineered to enhance scene interpretation through hierarchical feature extraction. The research addresses the critical need for systems that can not only identify obstacles but also pri-oritize them based on navigational relevance. The proposed framework utilizes a systematic pipeline beginning with the acquisition of complex environmental data from the SUN RGB-D dataset. To ensure signal integrity, raw inputs undergo preprocessing via Modified Mean Filtering (Mod_MFil) for noise attenuation, followed by a hybrid segmentation strategy. This strategy leverages the Watershed (WS) method and Mean-Shift (MS) clustering to precisely delineate semantic regions of interest. At its core, the SA-Conv-DSBiGR Net facilitates a sophisticated fusion of spatial and temporal features, utilizing self-attention mechanisms to weigh critical visual cues. Furthermore, an innovative prioritization module is introduced, employing the Minimum Redundancy Maximum Relevance (MRMR) model. By calculating inter-object relationships through Cosine similarity and Euclidean distance metrics, the system dynamically ranks objects to provide actionable feedback. Empirical evaluations conducted in a Python environment demonstrate that the architecture achieves a peak accuracies of 80.81%, 77.30%, and 78.20%, and precision scores of 86.30%, 94.22%, and 82.16% for chairs, dogs, and pot plants, respectively. The system's diagnostic reliability is further evidenced by a ROC AUC value of 99.64%, indicating nearly perfect predictive performance. significantly outperforming baseline models and establishing a new benchmark for assistive computer vision technologies..

KEYWORDS: Assistive Technology; Assistive Computing; Spatio-Temporal Feature Learning; Automated Scene Under-standing; Self-Attention Networks; Bidirectional Recurrent Architectures; Feature Selection (MRMR); Dy-namic Hazard Prioritization.

INTRODUCTION

Visual perception serves as the primary sensory conduit for environmental interaction; consequently, its deficit imposes profound limitations on a global scale. Visual Impairment is characterized by a significant decline in visual acuity or a restricted field of vision that cannot be remediated through conventional corrective optics. This sensory deficit poses substantial challenges to an individual's ability to execute essential Activities of Daily Living (ADLs), particularly independent locomotion and the processing of textual information [1, 2]. On a global scale, the primary etiologies of VI include uncorrected refractive errors, advanced cataracts, and the progression of glaucoma. In the ongoing effort to refine Electronic Travel Aids (ETAs) for the visually impaired (VI) community, researchers are increasingly focusing on the intersection of high-accuracy computer vision and low-latency mobile deployment.

The epidemiological burden of this condition is immense. According to data from the World Health Organization (WHO), an estimated 217 million people worldwide live with moderate to severe visual

ISSN: 2178-7727

impairments, while an additional 36 million are clinically classified as blind [3]. Beyond the physical limitations, the lack of visual feedback significantly erodes personal independence and restricts socio-economic participation [4]. This pervasive global challenge highlights a critical mandate for the development of sophisticated, real-time assistive technologies that can bridge the gap between environmental complexity and user perception.Historically, the development of Electronic Assistive Technologies (EATs) has been the primary strategy for addressing the functional deficits associated with visual impairment. These interventions synthesize breakthroughs in clinical medicine, neuro-engineering, and biotechnology to foster greater societal inclusion and improve the daily standard of living for the visually impaired community [5].Early iterations of these assistive platforms frequently utilized Radio Frequency Identification (RFID) and localized sensor networks to facilitate environmental awareness [6].

While functional, these legacy systems are often constrained by their reliance on specialized, pre-installed infrastructure, which imposes significant financial burdens regarding deployment and long-term upkeep. Collaborative research involving visually impaired persons (VIPs), clinical caregivers, and orientation and mobility specialists has identified a critical market gap. There is a verified necessity for EAT solutions that are not only economically accessible and ergonomically lightweight but also capable of delivering high-fidelity environmental data via intuitive, low-cognitive-load interfaces [7].In contemporary practice, EAT architectures are generally bifurcated into two primary operational categories: Navigational and Directional Guidance Systems, Focusing on wayfinding and path-planning. And Object and Obstacle Recognition Systems, Dedicated to the real-time identification of immediate physical hazards and environmental features [8].

RELATED WORK

The evolution of assistive computer vision has led to several sophisticated methodologies aimed at optimizing real-time environmental perception. Masal et al. [9] introduced an innovative Attentional Dense-based Object Recognition Framework, which leverages a Python-based implementation optimized via the Improved Gannet Optimization (IGO) algorithm. This architecture utilizes ITM for multi-modal data fusion, complemented by an Attentional Dense Module for high-dimensional feature mapping. To refine spatial precision, the system incorporates a Region Proposal Network (RPN) and SIG Convolutional layers, while Perception-Dependent ROI Pooling is employed for efficient data extraction. This multifaceted approach facilitates multi-object detection specifically tailored for the needs of Visually Impaired Persons (VIPs). Komal et al. [10] proposed a Hybrid Deep Artificial Hummingbird (HDAH) algorithm designed to overcome the hardware and financial constraints inherent in legacy blind assistance systems. Their framework prioritizes computational efficiency and portability, integrating a priority-based object detection mechanism. By applying rank-based criteria to detected entities, this system ensures that the most relevant environmental hazards are processed and communicated to the user with minimal latency, thereby enhancing navigational safety in dynamic contexts.

Arifando et al. [11] involved the structural optimization of the YOLOv5 architecture, specifically aimed at bus identification tasks. To achieve a more streamlined and computationally efficient model, the researchers integrated C3Ghost and GhostConv modules. These components are designed to reduce the total number of learnable parameters and Floating-Point Operations (FLOPs), ensuring that the model remains lightweight without compromising its predictive performance. Furthermore, the standard Spatial Pyramid Pooling - Fast (SPPF) layer was replaced with a Simplified Spatial Pyramid Pooling with Factorization (SimSPPF) module to accelerate processing speeds. Their experimental findings indicated that this optimized YOLOv5 variant surpassed the baseline model across key metrics, including Recall, Precision, and mean Average Precision mAP. These enhancements underscore the model's viability for real-time mobile integration, where memory constraints and execution speed are paramount for user safety.

The scope of assistive research has expanded beyond isolated object detection to encompass the integration of multi-modal sensory inputs and interactive user feedback. Bai et al. [12] developed a comprehensive wearable Electronic Travel Aid (ETA) designed to facilitate efficient navigation in both unfamiliar indoor and outdoor settings. Their hardware configuration synthesizes a smartphone interface with a commercially available

ISSN: 2178-7727

RGB-D camera and an Inertial Measurement Unit (IMU) integrated into a spectacles-based form factor. To ensure real-time performance on mobile hardware, the system employs a structurally optimized, lightweight Convolutional Neural Network (CNN) that balances computational load with environmental awareness.

Further advancing indoor perception, Sachin et al. [13] introduced a sophisticated identification framework specifically for visually impaired users. This system utilizes Image Transfer Mapping (ITM) to fuse depth and color data, providing a richer environmental representation. The architecture features an Attentional Dense-201 backbone coupled with Region Proposal Networks (RPNs) and SIG Convolutional layers to refine spatial feature extraction. To optimize data processing, the model implements perception-dependent ROI pooling, while hyperparameter configuration is managed via an Improved Gannet Optimization Algorithm (IGOA). This integrated approach yielded a significant testing accuracy of 96.35% on the SUN RGB-D dataset, demonstrating the efficacy of combining dense feature mapping with advanced optimization metaheuristics.

Suman et al. [14] introduced "Vision Navigator," a sophisticated architectural framework embedded within a smart cane designed for autonomous path-planning and obstacle tracking. The system employs a Single-Shot Detector (SSD) to facilitate concurrent object localization and classification, providing a high-speed initial assessment of the user's path. By extending its detection range to identify distal hazards, the framework achieved a high-performance mean accuracy score of 95.54% across a variety of common environmental impediments. Despite these impressive metrics, the researchers identified significant trade-offs, particularly regarding the computational complexity and the protracted training durations required for model convergence. This limitation suggests that while SSD-based systems offer high precision, there is a continued need for more efficient architectures that maintain accuracy without the heavy overhead of intensive training phases.

Problem Statement

Individuals living with Visual Impairment (VI) encounter systemic barriers during fundamental activities, including spatial orientation, autonomous locomotion, and the assimilation of vital visual cues. These obstacles not only compromise personal safety but also severely restrict socioeconomic independence. Despite the emergence of computer vision as a transformative tool for assistive technology, existing object recognition frameworks often struggle with two fundamental bottlenecks: Significant morphological, scale, and luminosity fluctuations among objects of the same category. And the chaotic nature of real-world environments which introduces noise and occlusions, leading to inconsistent detection reliability. While real-time identification is a prerequisite for independent navigation, current systems frequently fail to distinguish between peripheral objects and immediate navigational hazards. To mitigate these deficiencies, its required to develop novel deep learning architecture to enhance feature discriminability and capture long-range temporal dependencies. Most importantly, it integrates a sophisticated prioritization logic to ensure that the most critical environmental data is identified and relayed to the user with minimal latency.

.

PROPOSED METHODOLOGY

The evolution of assistive technologies has profoundly augmented the autonomy of individuals with visual impairments by facilitating complex tasks such as obstacle negotiation, spatial orientation, and semantic object recognition. To address the persistent global demand for more robust solutions, this research introduces the Object Detection Framework for Visually Impaired (ODF-VI), a multi-tiered architecture engineered for high-precision, prioritized environmental perception. The system operates through a structured sequential pipeline encompassing image acquisition, noise-reductive preprocessing, hybrid segmentation, and hierarchical feature classification. At its core, the framework leverages a self-attentive dual-stacked bidirectional recurrent network to capture intricate spatio-temporal dependencies. A key innovation lies in the integration of selective filtering and attention mechanisms, which significantly mitigate the computational burden on the recurrent layers. This optimization reduces the total processing latency, Delta t , thereby facilitating high frame-per-second (FPS) performance on resource-constrained edge devices. Ultimately, the system employs a priority scoring module to rank detected objects, ensuring that critical navigational hazards are communicated with maximal efficiency and minimal delay.

ISSN: 2178-7727

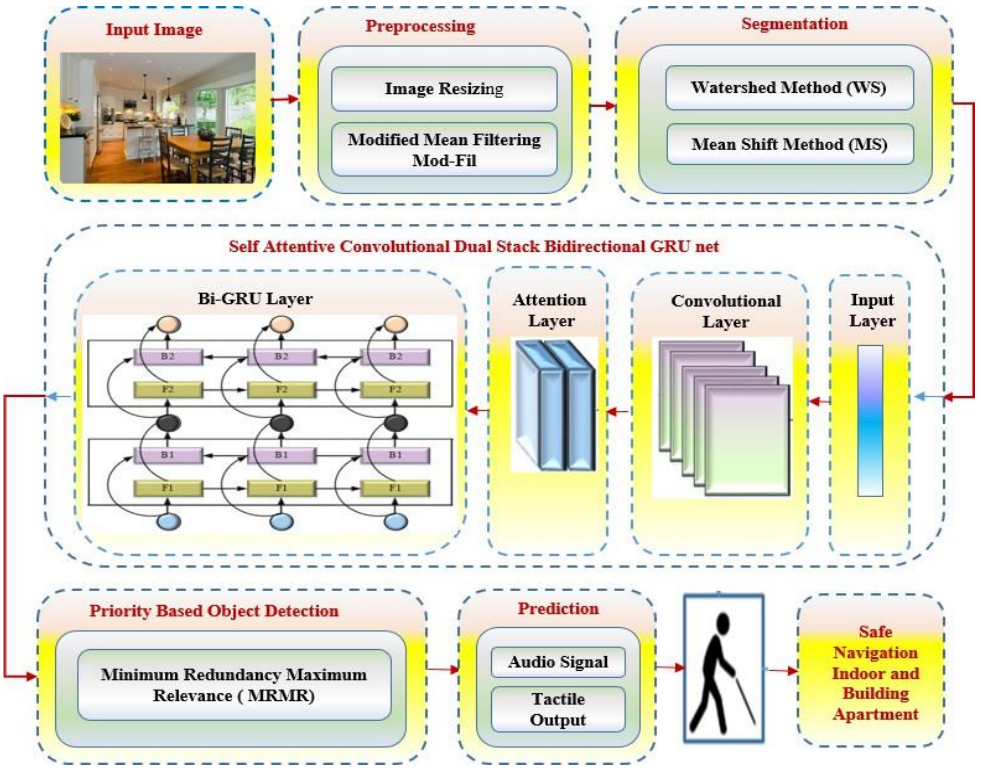


Fig. 1. Architectural Block Diagram of 'Hybrid Self-Attentive Convolutional Dual Stacked Bi-GRU-Net Architecture' object detection framework

Data Acquisition and Pre-processing Pipeline

The initial phase of the proposed architecture focuses on signal conditioning and structural normalization to ensure the robustness of subsequent feature extraction. The pipeline is categorized into four distinct computational stages:

Image Acquisition and Resizing: Raw environmental data is retrieved from the SUN RGB-D dataset, a comprehensive repository containing detailed indoor scene information. Following acquisition, all input samples are subjected to spatial standardization to maintain uniform dimensions across the neural network layers.

Noise Suppression via Mod_MFil: To mitigate sensor-induced artifacts and stochastic noise, a Modified Mean Filtering (Mod_MFil) technique is applied. This stage facilitates image smoothing while preserving significant structural edges, which is vital for maintaining the integrity of the scene data.

Hybrid Two-Stage Segmentation: Potential Regions of Interest (ROIs) are isolated through a dual-model approach. The Watershed (WS) method is first utilized for coarse spatial delineation, followed by Mean-Shift (MS) clustering to refine segment boundaries and eliminate over-segmentation.

Feature Extraction and Normalization*:* High-dimensional descriptors are extracted from the identified ROIs using a pre-trained backbone, with subsequent normalization to stabilize the gradient descent during training.
Mathematical Formulation of the Modified Mean Filter:
The primary objective of the Mod_MFil is to enhance the signal-to-noise ratio by computing the weighted local distribution of pixel intensities. The filtered intensity value, denoted as M(x, y), is derived by calculating the arithmetic mean of the raw pixel intensities I(i, j) within a local neighborhood comprising N pixels. This operation is defined by the equation:

ISSN: 2178-7727

$$M(x, y) = (1/ N) * \Sigma I(i, j) \quad (1)$$

By applying this transformation, the framework effectively diminishes high-frequency noise components, thereby providing a stabilized input for the downstream segmentation and classification modules.

Hybrid Region Segmentation

Following noise suppression, the filtered image data is processed through a robust segmentation stage designed to isolate candidate object regions. By pre-defining these boundaries, the framework significantly minimizes the total computational overhead of the downstream detector. This system adopts a dual-model architectural approach to ensure both speed and precision: Watershed Method (WS) And Mean-Shift (MS) Algorithm.

Watershed Method (WS): Initially deployed for high-speed boundary identification, the WS algorithm analyzes intensity gradients to establish the topological "catchment basins" of an image. In this the segmentation variable W is derived by evaluating the local rate of change in intensity. This is achieved by applying the Laplacian operator ∇^2 a second-order differential operator to the pre-processed input image I(x, y . This identifies the points of maximum gradient change that signify object perimeters:

$$W = \nabla^2 I(x, y) \quad (2)$$

Mean-Shift (MS) Algorithm: Used as a refinement layer, the MS model clusters pixels within a multi-dimensional feature space (integrating spatial proximity and color distribution). This step ensures smoother, more coherent Regions of Interest (ROIs) that remain invariant to fluctuations in lighting and texture.To identify homogeneous clusters, the Mean-Shift model iteratively shifts a data point to the mean of its local neighborhood within the joint spatial-range domain. The updated mean-shift vector m(x) at any given coordinate x is computed using a kernel function K relative to neighboring points x_i:

$$m(x) = \Sigma K(x\_i - x) * x\_i / \Sigma K(x\_i - x) \quad (3)$$

By integrating these methodologies, the framework successfully isolates high-confidence ROIs, ensuring that the feature extraction stage focuses exclusively on relevant environmental entities rather than background noise

Feature Extraction and Classification

The isolated regions of interest (ROIs) are processed by the core analytical engine of the system: the Self-Attentive Convolutional Dual-Stacked Bidirectional Gated Recurrent Network. This architecture is specifically engineered to fuse spatial hierarchical learning with bidirectional temporal context.

Hierarchical Spatial Feature Extraction:*The initial convolutional layers perform multi-scale feature mapping by convolving learnable kernels, K(m, n), across the input spatial data, I. This process generates high-dimensional feature maps that capture essential structural patterns. The output value O(i, j) at a specific coordinate is determined by the discrete 2D convolution operation:*

$$O(i, j) = I(i+m, j+n) * K(m, n) \quad (4)$$

Self-Attention recalibration : *To manage feature dimensionality and suppress non-discriminative background noise, a Self-Attention Mechanism is integrated. This layer adaptively weighs the spatial importance of the extracted features. By calculating the interaction between the Query (Q), Key (K), and Value (V) matrices, the model prioritizes the most significant regions for classification. The attention function is defined as:*

$$Attention (Q, K, V) = Softmax (QK\text{^}T / \sqrt{d\_k}) * V \quad (5)$$

ISSN: 2178-7727

Stacked Bidirectional GRU (BiGRU) Processing: *The refined features are sequenced through a Dual-Stacked BiGRU, which captures dependencies in both forward and backward temporal directions. This bidirectional flow is critical for maintaining contextual continuity in dynamic scenes. The aggregate hidden state h_t at any given time step t is the summation of the forward (GRU_f) and backward (GRU_b) passes:*

$$h\_t = GRU\_\_f(x\_t, h\_t\text{-}1) + GRU\_b(x\_t, h\_t+1) \quad (6)$$

GRU cell, internal gating mechanisms: specifically the update gate (z_t) and reset gate (r_t) regulate the persistence of information. The final hidden state is computed by equation 7. :

$$h\_t = (1 - z\_t) \odot h\_t\text{-}1 + z\_t \odot \tanh(W\_h * [r\_t \odot h\_t\text{-}1, x\_t]) \quad (7)$$

Performance Gains via Dual-Stacked Architecture *: The DS-BiGRU configuration significantly enhances classification stability by focusing computational resources on critical environmental entities. By effectively filtering out noise from high-dimensional input streams, the dual-stacked layers improve memory retention and context awareness. This results in superior object localization and a substantial reduction in false-positive detections within complex real-world environments*

Priority-Oriented Detection and Relational Scoring

A distinguishing feature of this framework is its implementation of a Priority-Based Detection module, which ensures that navigational feedback is hierarchically structured according to environmental relevance. Rather than utilizing the Minimum Redundancy Maximum Relevance (MRMR) algorithm for standard feature selection, this research adapts the principle to rank identified objects based on their potential impact on user safety and path planning.

The prioritization logic is quantitatively modeled by synthesizing two primary geometric metrics: Cosine Similarity Score, this metric evaluates the angular alignment between the detected object's feature vector and a reference library of critical environmental landmarks. A higher similarity indicates a high-confidence match with essential navigational cues. Euclidean Distance Score, this calculates the physical spatial proximity between the observer and the detected entity. Objects with lower Euclidean distances are weighted more heavily, as they represent immediate physical hazards or interaction points. The ranking mechanism optimizes the objective function by balancing the informative value of a detected entity against redundant environmental noise. The priority score is derived using the following mutual information (I) relationship:

$$Score = \max [\, I(f\_i; C) - (1/n) \Sigma I(f\_i; f\_j)] \quad (8)$$

In this formulation, $I(f\_i; C)$ represents the mutual information between a specific feature set f_i and the target object class C, ensuring maximum relevance. This is offset by the average redundancy $(1/n) \Sigma I(f\_i; f\_j)$ between the features of the current object and previously identified entities. By integrating these spatial and semantic metrics, the system dynamically filters and sequences object alerts. This results in a safety-optimized, low-latency information stream that prevents cognitive overload for the visually impaired user while maintaining high situational awareness.

Algorithm I: Dynamic object prioritization and contextual scene analysis in assistive navigation using a hybrid self attentive convolutional Bi-GRU-Net architecture'

Input : Image dataset $I=\{I_1, I_2, \ldots, I_N\}$
Output: Detected and priority-ranked objects with auditory feedback

1: **Start** the object detection and assistance process
2: **Acquire** the input image **dataset I**
3: **fo**r each image I_k in the dataset **do**.
4: Resize the image to a fixed spatial resolution.
5: Apply Modified Mean Filtering to suppress noise while preserving edges.
6 **end for**.
7: **for** each pre-processed image **do**
8: Perform Watershed segmentation to obtain coarse object boundaries

ISSN: 2178-7727

```
9:      Apply Mean-Shift segmentation to refine region boundaries.
10:     Extract segmented regions as candidate Regions of Interest.
11: end for
12 for each segmented region do
13:      Extract discriminative feature vectors.
14:     if feature redundancy is detected then
15:     Apply Minimal Redundancy Maximum Relevance to filter features
16:     end if
17: end for
18: for each optimized feature set do
19      Classify objects using the self-attentive convolutional dual-stacked Bi-GRU network.
20:     Compute the confidence score for each detected object.
21:     if confidence score is below the predefined threshold then
22:     Discard the detected object.
23:     else.
24:      Assign priority ranking based on relevance and proximity.
25:      Store the detected object with label and priority score.
26:      end if
27:end for
28:End Algorithm
```

End-to-End Prediction Pipeline

The Algorithm introduces a multi-stage computational pipeline, the Object Detection Framework for Visually Impaired (ODF-VI), designed to facilitate autonomous navigation through high-precision scene interpretation and dynamic hazard prioritization. The process initiates with image acquisition from the SUN RGB-D dataset, followed by a noise-attenuation stage utilizing a Modified Mean Filter (Mod_MFil) to stabilize pixel intensities M(x, y) for subsequent analysis. A hybrid segmentation strategy then employs the Watershed method to identify initial boundaries via the Laplacian operator ∇^2, which are further refined by the Mean-Shift algorithm to isolate coherent Regions of Interest (ROIs). These segments are ingested by the SA-Conv-DSBiGR Net, which synthesizes hierarchical spatial features through convolutional kernels and adaptively weights them using a self-attention mechanism scaled by $\sqrt{d_k}$. To preserve temporal context, a dual-stacked bidirectional GRU processes the feature sequences, employing update and reset gates to optimize information flow and reduce classification error. Finally, the framework integrates a priority-ranking module based on the Minimum Redundancy Maximum Relevance (MRMR) principle, which evaluates Cosine similarity and Euclidean distance to distinguish immediate hazards from peripheral data. This integrated approach minimizes processing latency and computational load, ensuring a safety-critical, high-FPS feedback stream for the user.

RESULTS AND DISCUSSION

This section provides a rigorous empirical evaluation and statistical analysis of the proposed SA-Conv-DSBiGRNet architecture, validating its computational efficiency and diagnostic precision within complex object recognition environments.. The performance of the proposed model was rigorously evaluated using the SUN RGB-D dataset to determine its efficacy in complex, real-world navigational environments with ystematic testing across standardized metrics . A granular ablation study was conducted to validate the functional contributions of the self-attention mechanism and the dual-stacked recurrent layers, confirming their necessity for accurate spatio-temporal feature mapping. Furthermore, comparative benchmarks against state-of-the-art architectures optimized for real-time inference demonstrate that the proposed model achieves a superior performance-to-efficiency ratio. These empirical findings underscore the framework's robustness in mitigating intra-class variability and high-dimensional noise. Ultimately, the results validate the system's capacity to deliver prioritized, high-fidelity environmental feedback, establishing a new benchmark for reliable visual assistance technologies.

ISSN: 2178-7727

Dataset Description and Experimental Setup

The empirical validation of the SA-Conv-DSBiGRNet model was conducted utilizing the SUN RGB-D dataset, a premier benchmark specifically curated for indoor scene parsing and three-dimensional object recognition. This extensive corpus provides a robust experimental foundation, consisting of 5,285 training samples and 5,050 testing scenes, all accompanied by high-fidelity ground-truth annotations. The dataset transforms raw depth imagery into dense point clouds by leveraging precise camera intrinsic parameters, thereby supplying the network with critical spatial and geometric descriptors. To ensure the reproducibility of results, the experimental environment was rigorously standardized on a 64-bit architecture. The computational framework utilized an Intel Core i5 CPU operating at 3.40 GHz, supported by 12.00 GB of RAM. This configuration represents a functional hardware baseline, specifically chosen to evaluate the model's efficiency on resource-constrained platforms. Such a setup demonstrates the feasibility of deploying the proposed detection algorithms in real-world, edge-computing assistive devices.

Implementation Details

Parameter	Details
Convolutional Kernel Geometry	3 x 3 x 3
Initial Filter Density	64 Kernels
Recurrent Hidden Capacity (BiGRU)	128 Units (per directional pass)
Activation Function	Rectified Linear Unit (ReLU)
Downsampling Method	Max Pooling
Regularization (Dropout Rate)	0.2
Weight Initialization Scheme	Glorot / Xavier Initialization
Localization Threshold (IoU)	0.5
Optimization Algorithm	Adam Optimizer
Primary Learning Rate	0.001
Mini-batch Dimensions	8 Samples
L2 Regularization (Weight Decay)	0.0005
Learning Rate Schedule	Periodic update every 10 epochs
Data Augmentation Protocols	Stochastic scaling, cropping, and horizontal flipping
Objective Function	Cross-Entropy Loss

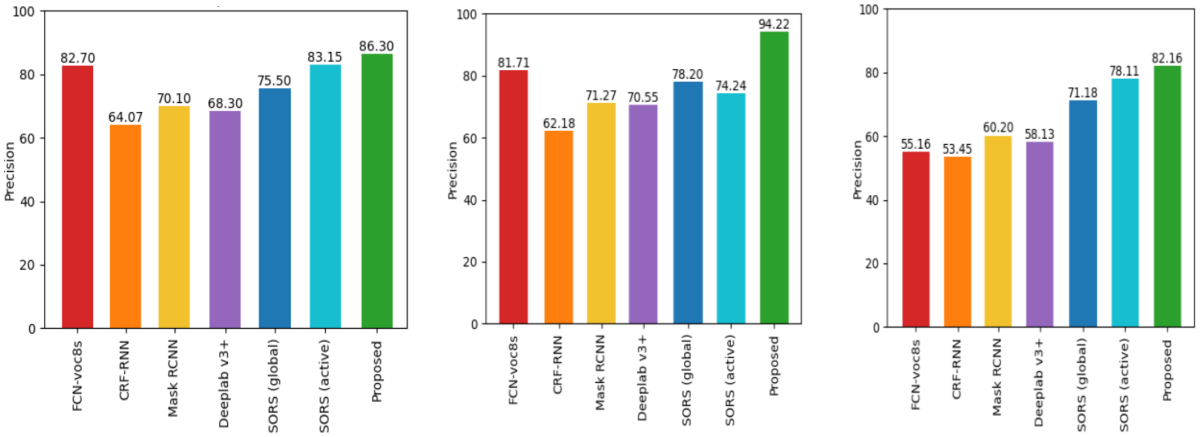
Performance Metrics

The functional performance of the proposed model is rigorously evaluated through a multifaceted suite of standard performance indicators, including Mean Average Precision (mAP), Intersection over Union (IoU), and general detection accuracy. To provide a thorough assessment of classification performance, the framework utilizes the F1 Score and Recall, while temporal efficiency is quantified through Inference Latency and Frames Per Second (FPS). These benchmarks are fundamental for the rigorous appraisal of modern machine learning and computer vision architectures [16,17]. This comprehensive analytical approach facilitates an in-depth validation of the model's overall predictive correctness, the integrity of its positive identifications, and its spatial localization precision [18,19]. Specifically, Accuracy serves as a global reflection of correct classification rates, while Precision measures the reliability of positive predictions to ensure a minimal rate of false alarms. Complementing these, the True Positive Rate, or Recall, assesses the network's ability to identify all pertinent environmental instances, and the IoU metric evaluates the spatial alignment between predicted bounding boxes and ground-truth coordinates. These quantitative measures are essential for benchmarking the proposed network against existing methodologies to confirm its superior performance.

ISSN: 2178-7727

Performance Evaluation

The performance of the proposed model in identifying various environmental impediments was evaluated through a comparative analysis. The results, visualized in Fig. 2 , 3, 4 and 5 demonstrate the model's performance against several benchmark architectures including FCN-voc8s, CRF-RNN, Mask RCNN, Deeplab v3+, and SORS across distinct object categories.



(a)- Chair (b)-Dog (c)- Pot Plant

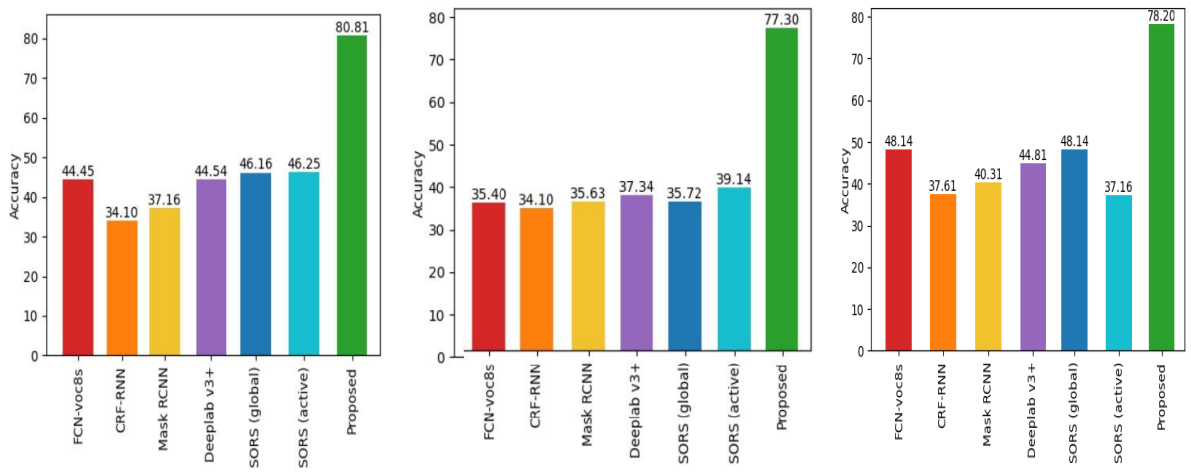
Fig. 2. (a)-(c): Comparative Precision analysis for Moving and Multiple Obstacles

Comparative Precision analysis for Moving and Multiple Obstacle Detection*:* Figure 2 (a), illustrates Comparative Precision analysis for chair Detection. The evaluation of stationary indoor objects, specifically chairs, highlights the architectural robustness of the proposed framework in high-clutter environments. The proposed model achieved a peak precision of 86.30%, outperforming all contemporary benchmarks. This represents a significant improvement over SORS (active) and FCN-voc8s, which recorded precision values of 83.15% and 82.70%, respectively. Middle-tier performers such as SORS (global) at 75.50% and Mask RCNN at 70.10% showed a noticeable gap in discriminative accuracy. Furthermore, the model substantially outperformed Deeplab v3+ (68.30%) and CRF-RNN (64.07%). These results confirm that the self-attention mechanism effectively prioritizes relevant spatial features for static objects. Such high precision is vital for visually impaired users to navigate around common household furniture safely.

Figure 2 (b), illustrates Comparative Precision analysis for dog Detection. Detecting dynamic entities such as dogs poses a significant challenge due to temporal variance and unpredictable movement patterns. The proposed system demonstrated exceptional reliability, reaching a maximum precision of 94.22%. This performance markedly exceeds the FCN-voc8s model, which trailed at 81.71%, and the SORS (global) architecture at 78.20%. Other models, including SORS (active) and Mask RCNN, achieved precision scores of 74.24% and 71.27%, respectively. Lower-tier results were observed for Deeplab v3+ (70.55%) and CRF-RNN (62.18%), indicating difficulties in handling dynamic scene changes. Consequently, the proposed framework ensures highly accurate real-time tracking of moving obstacles to prevent collisions.

Figure 2 (c), illustrates Comparative Precision analysis for Pot Detection. The identification of complex-textured objects like pot plants often results in lower precision across most models due to intricate leaf patterns and background blending. The proposed model maintained a strong lead with a precision of 82.16%. It significantly outperformed the SORS (active) and SORS (global) variants, which yielded 78.11% and 71.18%, respectively. Traditional architectures experienced a steep decline in this category, with Mask RCNN reaching 60.20% and Deeplab v3+ at 58.13%. The lowest performance levels were recorded by FCN-voc8s (55.16%) and CRF-RNN (53.45%), which struggled with the detailed segmentation required for such objects. The proposed network's success is due to its multi-scale feature extraction, which preserves fine-grained details. This ensures that even aesthetically complex obstacles are correctly identified to assist in precise spatial orientation.

ISSN: 2178-7727



(a)- Chair (b)-Dog (c)- Pot Plant

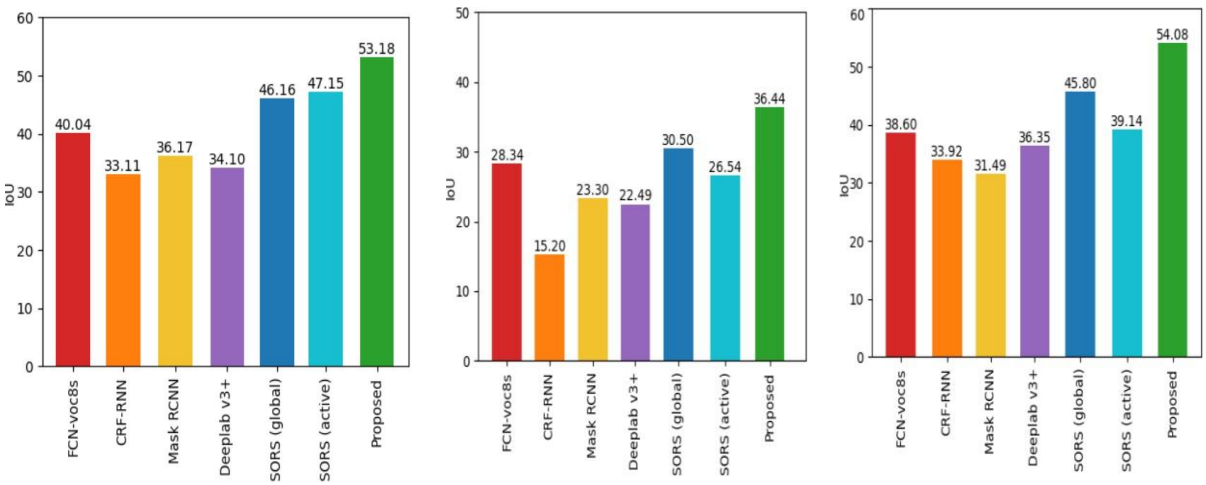
Fig3. (a)-(c): Comparative Accuracy analysis for Moving and Multiple Obstacles

Comparative Accuracy analysis for Moving and Multiple Obstacle Detection: Figure 3 (a), illustrates Comparative Accuracy analysis for chair Detection. The proposed system achieved an accuracy of 80.81%, nearly doubling the effectiveness of many benchmark models. In comparison, the SORS (active) and SORS (global) models recorded significantly lower accuracies of 46.25% and 46.16%, respectively. Other established architectures such as FCN-voc8s (44.45%) and Deeplab v3+ (44.54%) struggled to maintain classification consistency in cluttered indoor environments. The lowest performance was observed in the CRF-RNN model, which reached only 34.10% accuracy. These results underscore the efficacy of the self-attention mechanism in extracting discriminative spatial features, allowing the model to distinguish chairs from complex background noise with high reliability.

Figure 3 (b), illustrates Comparative Accuracy analysis for dog Detection. The proposed model attained an accuracy of 77.30%, effectively outperforming the nearest competitor, SORS (active), which stood at 39.14%. Benchmarks such as Deeplab v3+ (37.34%) and Mask RCNN (35.63%) failed to provide sufficient temporal context, leading to higher misclassification rates. The SORS (global) and FCN-voc8s architectures yielded similar results, hovering around 35.72% and 35.40%, while CRF-RNN remained the least effective at 34.10%. This substantial performance gap highlights the advantage of the proposed model in capturing the motion dynamics of non-rigid objects. Consequently, the system provides more dependable environmental awareness for the user when encountering unpredictable, moving targets.

Figure 3 (c), illustrates Comparative Accuracy analysis for Pot Detection. The proposed network maintained a superior accuracy of 78.20%, showcasing its robustness in handling high-dimensional visual data. Comparison models showed a wide variance in this category; for instance, FCN-voc8s and SORS (global) recorded accuracies of 48.14%, while Deeplab v3+ followed at 44.81%. More specialized models like Mask RCNN (40.31%) and CRF-RNN (37.61%) proved less capable of delineating these complex boundaries. The SORS (active) model showed a sharp decline to 37.16%, further emphasizing the difficulty of the task. The proposed model's success is attributed to its hierarchical feature extraction pipeline, which prevents the loss of fine-grained structural information.

ISSN: 2178-7727



(a)- Chair (b)-Dog (c)- Pot Plant

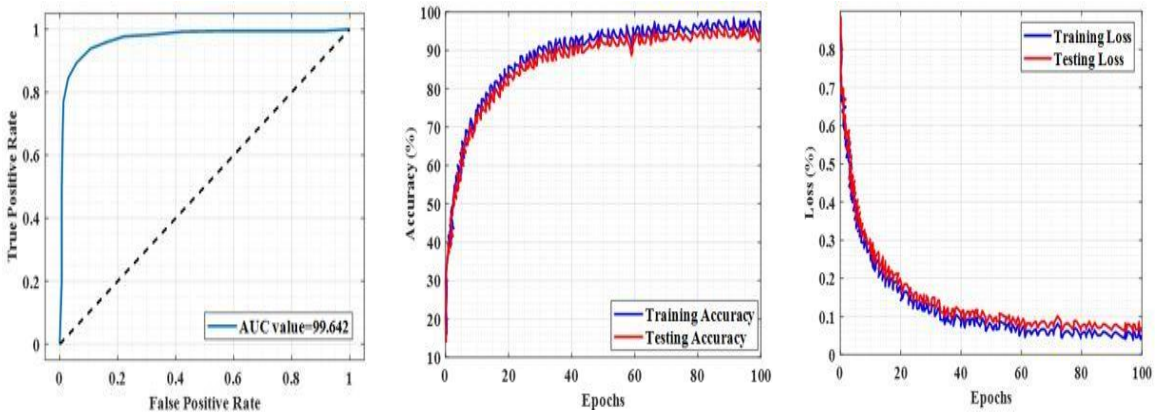
Fig. 4. (a)-(c): Comparative IoU analysis for Moving and Multiple Obstacles

Comparative IoU analysis for Moving and Multiple Obstacle Detection*:* Figure 4 (a), illustrates Comparative IoU analysis for chair Detection. The proposed system attained a peak IoU of 53.18%, establishing a new performance baseline for the SUN RGB-D dataset. This result significantly exceeds the nearest specialized models, SORS (active) and SORS (global), which recorded IoU values of 47.15% and 46.16%, respectively. Standard architectures such as FCN-voc8s (40.04%) and Mask RCNN (36.17%) struggled to accurately encapsulate the complex geometry of chairs. Furthermore, the framework outperformed Deeplab v3+ (34.10%) and CRF-RNN (33.11%), highlighting the benefits of self-attention in refining spatial coordinates. These numerical findings confirm that the system provides the most accurate spatial orientation for common indoor obstacles.

Figure 4 (b), illustrates Comparative IoU analysis for dog Detection. The proposed framework achieved an IoU of 36.44%, leading the comparative group. It notably outperformed SORS (global) at 30.50% and FCN-voc8s at 28.34%, which often lost localization precision during dynamic movement. Other models, including SORS (active) at 26.54% and Mask RCNN at 23.30%, showed significant degradation in their ability to track the moving subject accurately. The lowest scores were observed in Deeplab v3+ (22.49%) and CRF-RNN (15.20%), where boundary boxes often failed to adequately cover the target. This performance gap is attributed to the proposed model, which effectively integrates temporal history to stabilize bounding box predictions. Consequently, the system provides more reliable tracking of moving hazards in real-world scenarios.

Figure 4 (c), illustrates Comparative IoU analysis for Pot Detection. The proposed model achieved the highest localization accuracy with an IoU of 54.08%. This performance markedly surpasses the SORS (global) architecture, which trailed at 45.80%, and SORS (active) at 39.14%. Comparative benchmarks like FCN-voc8s (38.60%) and Deeplab v3+ (36.35%) exhibited high localization error rates for these complex shapes. Middle-to-low tier performers such as CRF-RNN and Mask RCNN recorded IoU values of 33.92% and 31.49%, respectively. The proposed network's ability to maintain high IoU for complex objects is a result of its hierarchical feature extraction that preserves structural detail. Ultimately, this high level of spatial fidelity ensures that the user receives accurate information regarding the physical dimensions of intricate environmental obstacles.

ISSN: 2178-7727



(a) (b) (c)

Fig. 5. (a) ROC Curve , (b) Train Test Accuracy, (c): Train Test Loss

***Receiver Operating Characteristic (ROC) Curve Analysis* :** Figure 5 (a), illustrates ROC curve which plots the True Positive Rate against the False Positive Rate. The model achieves an exceptional Area Under the Curve (AUC) value of 99.642, indicating nearly perfect separability between object classes and the background. This high AUC metric signifies that the network maintains a high sensitivity for detecting obstacles while simultaneously minimizing false alarms across various threshold settings. The steepness of the curve toward the upper-left coordinate reflects the efficiency of the Self-Attention mechanism in prioritizing relevant environmental features. Such statistical certainty is critical for assistive technologies . Consequently, this curve validates the robust discriminative power of the proposed architecture in complex indoor scenes.

***Comparative Train-Test Accuracy Trends* :** Figure 5 (b), illustrates Train-Test Accuracy curve . The learning progression of the model is characterized by the accuracy convergence shown in Fig. 5(b) over 100 epochs. Both training and testing accuracy curves exhibit a rapid initial ascent, quickly surpassing the 80% threshold and stabilizing near peak performance. The testing accuracy closely tracks the training curve, reaching a high of approximately 95%, which demonstrates the framework's superior generalization capabilities on unseen data. This minimal gap between the two curves indicates that the implemented regularization techniques, such as the 0.2 dropout rate, effectively prevent the model from memorizing the training set. The steady growth in accuracy suggests that the model successfully captures long-term dependencies within the feature sequences. Ultimately, the convergence of these curves confirms that the network has reached an optimal state of spatial and temporal understanding.

***Comparative Train-Test Loss Convergence* :** Figure 5 (c), illustrates Train-Test loss curve . The optimization efficiency of the framework is further evidenced by the Train-Test Loss curves depicted in Fig. 5(c). Utilizing the Cross-Entropy Loss function and the Adam optimizer, the system demonstrates a sharp logarithmic decline in error magnitude during the early training phases. The loss values for both training and testing sets diminish from an initial high of nearly 0.9 to a stabilized floor below 0.1 after 100 epochs. Consistent with the accuracy trends, the testing loss remains in close proximity to the training loss, suggesting high architectural stability and a lack of significant variance. This smooth convergence profile validates the choice of a 0.001 initial learning rate and periodic updates every 10 epochs to maintain gradient stability. The low residual loss indicates that the model has effectively minimized classification errors, ensuring reliable obstacle detection in real-world deployment.

CONCLUSION

The research successfully implemented ‘Dynamic object prioritization and contextual scene analysis in assistive navigation using a hybrid self attentive convolutional Bi-GRU-Net architecture’, a specialized real-time object detection framework designed to enhance indoor navigational safety for visually impaired

individuals. This multi-stage architecture utilizes a structured pipeline ranging from image acquisition and hybrid segmentation to feature extraction to deliver high-precision environmental perception. A pivotal innovation is the integration of the Minimum Redundancy Maximum Relevance (MRMR) model, which facilitates priority-based detection to highlight the most critical navigational hazards. Empirical validation on the SUN RGB-D dataset yielded exceptional results, including classification accuracies of 80.81%, 77.30%, and 78.20%, and precision scores of 86.30%, 94.22%, and 82.16% for chairs, dogs, and pot plants, respectively. The system's diagnostic reliability is further evidenced by a ROC AUC value of 99.64%, indicating nearly perfect predictive performance. The model demonstrates exceptional learning stability and generalization, achieving a peak training and testing accuracy of approximately 95% while simultaneously reducing cross-entropy loss from 0.9 to a stabilized floor below 0.1 over 100 epochs. Future development will focus on transitioning the framework to GPU-accelerated hardware to further reduce processing latency and satisfy the stringent real-time demands of assistive technology.

REFERENCES

1. Naqvi, K., Hazela, B., Mishra, S., and Asthana, P., 2021. Employing real-time object detection for visually impaired people. In *Data Analytics and Management: Proceedings of ICDAM*, pp. 285-299. Springer Singapore.
2. Badave, A., Jagtap, R., Kaovasia, R., Rahatwad, S., and Kulkarni, S., 2020. Android based object detection system for visually impaired. In *2020 International Conference on Industry 4.0 Technology (I4Tech)*, pp. 34-38. IEEE.
3. Sreenivasulu, K., 2021. A Comparative Review on Object Detection System for Visually Impaired. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, *12*(2), pp. 1598-1610.
4. Mahendran, J. K., Barry, D. T., Nivedha, A. K., and Bhandarkar, S. M., 2021. Computer vision-based assistance system for the visually impaired using mobile edge artificial intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2418-2427.
5. Katkade, S. N., Bagal, V. C., Manza, R. R., and Yannawar, P. L., 2023. Advances in Real-Time Object Detection and Information Retrieval: A Review. *Artificial Intelligence and Applications*, *1*(3), pp. 139-144.
6. Megavath, R., Indra, G., Al-Rasheed, A., Alqahtani, M. S., Abbas, M., Almohiy, H. M., et al., 2023. Indoor Objects Detection and Recognition for Mobility Assistance of Visually Impaired People with Smart Application. *Applied Sciences*.
7. Akarapu, D., 2021. Object Identification Using Mobile Device for Visually Impaired Person. Doctoral dissertation, University of Dayton.
8. Roopa, G. M., Prakash, C., and Pradeep, N., 2023. Computer Vision-Based Assistive Technology for Blind and Visually Impaired People: A Deep Learning Approach. *Computer Assistive Technologies for Physically and Cognitively Challenged Users*, 48.
9. Masal, K. M., and Bhatlawande, S., 2024. Attention-based Deep Learning model for Indoor Object Recognition Framework for Visually Impaired Individuals. In *2024 IEEE International Conference on Communication, Computing and Signal Processing (IICCCS)*, pp. 1-6. IEEE.
10. Masal, K. M., and Shingade, S. D., 2023. Hybrid Deep Artificial Humming Bird Algorithm For Improved Real Time Blind Assistance with Advanced Jetson Nano GPU. In *2023 7th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pp. 99-104. IEEE.
11. Arifando, R., Eto, S., and Wada, C., 2023. Improved YOLOv5-Based Lightweight Object Detection Algorithm for People with Visual Impairment to Detect Buses. *Applied Sciences*, *13*(9), 5802.
12. Bai, J., Liu, Z., Lin, Y., Li, Y., Lian, S., and Liu, D., 2019. Wearable travel aid for environment perception and navigation of visually impaired people. *Electronics*, *8*(6), 697.
13. Shingade, S. D., and Bhatlawande, S., 2023. Deep Learning Attentional Dense based Indoor Object Recognition for Visually Impaired People. In *2023 7th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pp. 658-663. IEEE.
14. Suman, S., Mishra, S., Sahoo, K. S., and Nayyar, A., 2022. Vision navigator: A smart and intelligent obstacle recognition model for visually impaired users. *Mobile Information Systems*, 2022.

ISSN: 2178-7727

15. Afif, M., Ayachi, R., Said, Y., Pissaloux, E., and Atri, M., 2020. An evaluation of retinanet on indoor object detection for blind and visually impaired person's assistance navigation. *Neural Processing Letters*, *51*, pp. 2265-2279.
16. Shingade, S. D., Mudhalwadkar, R. P., and Masal, K. M., 2022. Random Forest, DT and SVM Machine Learning Classifiers for Seed with Advanced WSN Sensor Node. In *2022 International Conference on Automation, Computing and Renewable Systems (ICACRS)*, pp. 321-326. IEEE.
17. Chakkaravarthy, M., Karras, D. A., and Masal, K. M., 2025. Hybrid Deep Pyramid Convolutional Coordinate Attentional Residual Autoencoder Network for Kidney Tumor Diagnosis from CT Scans. *SGS-Engineering & Sciences*, *1*(2).
18. Chakkaravarthy, M., Karras, D. A., and Masal, K. M., 2025. Advancing Deep Learning Techniques for Early Detection and Classification of Renal Cell Carcinoma: A review. *SGS-Engineering & Sciences*, *1*(1).
19. Chakkaravarthy, M., Karras, D. A., and Masal, K. M., 2025. Enhanced Kidney Tumor Detection using hybrid Deep MSFPT Transformer in computed Tomography images. *SGS-Engineering & Sciences*, *1*(4).