RESEARCH ARTICLE

Advancing Robotic Automation with Custom Sequential Deep CNN-Based Indoor Scene Recognition

Fida Hussain Dahri[1], **Ghulam E Mustafa Abro**[2,*], **Nisar Ahmed Dahri**[3], **Asif Ali Laghari**[4] **and Zain Anwar Ali**[5]

[1] School of Computer Science and Engineering, Southeast University, Nanjing 211189, China
[2] Interdisciplinary Research Centre for Aviation and Space Exploration (IRC-ASE), King Fahd University of Petroleum and Minerals (KFUPM), Dhahran, 31261, Kingdom of Saudi Arabia
[3] Faculty of Social Sciences and Humanities, School of Education, University Technology Malaysia, Malaysia
[4] Software College, Shenyang Normal University, Shenyang 110136, China
[5] Electronic Engineering Department, Maynooth International Engineering College (MIEC), Maynooth University, Maynooth, Co. Kildare, Ireland

Abstract

Indoor scene recognition poses considerable hurdles, especially in cluttered and visually analogous settings. Although several current recognition systems perform well in outside settings, there is a distinct necessity for enhanced precision in inside scene detection, particularly for robotics and automation applications. This research presents a revolutionary deep Convolutional Neural Network (CNN) model tailored with bespoke parameters to improve indoor picture comprehension. Our proprietary dataset consists of seven unique interior scene types, and our deep CNN model is trained to attain excellent accuracy in classification tasks. The model exhibited exceptional performance, achieving a training accuracy of 99%, a testing accuracy of 89.73%, a precision of 90.11%, a recall of 89.73%, and an F1-score of 89.79%. These findings underscore the efficacy of our methodology in tackling the intricacies of indoor scene recognition. This research substantially advances the domain of robotics and automation by establishing a more resilient and dependable framework for autonomous navigation and scene comprehension in GPS-denied settings, facilitating the development of more efficient and intelligent robotic systems.

Keywords: indoor scene recognition, deep convolutional neural network (CNN), robotics and automation autonomous navigation and GPS-Denied environments.

1 Introduction

In recent years, robotic automation has gained Robotic automation has gained prominence in various sectors, such as manufacturing, logistics, healthcare, and domestic assistance [1, 2]. Implementing robots in interior settings, where they independently execute activities, has distinct obstacles, particularly in scene recognition [3, 4]. Indoor spaces, including residences, workplaces, medical facilities, and

Academic Editor:
Quanmin Zhu

Submitted: 17 September 2024
Accepted: 09 December 2024
Published: 27 December 2024

Vol. 2, **No.** 1, 2025.
10.62762/TIS.2025.613103

***Corresponding author:**
Ghulam E Mustafa Abro
mustafa.abro@ieee.org

Citation
Dahri, F. H., Abro, G. E. M., Dahri, N. A., Laghari, A. A., & Ali, Z. A. (2024). Advancing Robotic Automation with Custom Sequential Deep CNN-Based Indoor Scene Recognition. *ICCK Transactions on Intelligent Systematics*, 2(1), 14–26.

© 2024 ICCK (Institute of Central Computation and Knowledge)

industrial sites, encompass diverse objects, structures, and configurations [7–9], hindering the precise identification and understanding of the surroundings. The challenge stems from the resemblance of objects in various interior environments, such as chairs present in both auditoriums and computer labs, resulting in recognition problems [10–13]. The intricacy necessitates enhancing indoor scene identification in robotic automation [5, 6].

Indoor scene recognition poses significant challenges due to the visual similarity of objects across various environments, such as chairs in auditoriums and computer labs, which leads to recognition errors. While outdoor scene recognition algorithms have shown considerable success, their performance diminishes substantially in indoor settings. This limitation necessitates the development of a specialized deep-learning approach tailored for indoor environments. By leveraging a custom dataset, this research addresses these challenges. It demonstrates the effectiveness of the proposed Sequential Deep CNN model for improving accuracy and robustness in robotic indoor scene recognition.

Conventional rule-based methods and manually designed scene templates inadequately capture the complex visual details in many indoor environments [14, 15]. Enhancing scene recognition can substantially improve robotic automation by allowing robots to accurately detect, understand, and adapt to their environment [16, 17]. This skill enhances fundamental robotic operations, including object manipulation, path planning, and semantic comprehension [18, 19]. Mobile robots necessitate an elevated degree of semantic comprehension to execute intricate operations in dynamic indoor settings [20].

In light of these obstacles, effective indoor scene identification continues to be a significant concern in robotics [3, 21]. Robots must identify and adjust to changing environments to traverse and interact efficiently in GPS-denied contexts [22, 23]. Our study fulfils this requirement by creating a deep learning-based system customized for the particular needs of indoor spaces. We provide a bespoke dataset comprising seven unique indoor scene categories and propose a deep Convolutional Neural Network (CNN) model to extract pertinent characteristics and classify images with high precision. The model demonstrates substantial enhancements in classification, as seen by its elevated training and testing accuracy on the custom dataset.

Although outside scene recognition algorithms have achieved success [13, 24, 25], their efficacy diminishes considerably when utilized in inside environments [15]. Utilizing a bespoke dataset, our research illustrates the effectiveness of a deep learning strategy designed for indoor scene detection. The suggested technology enhances robotic automation by augmenting robots' capacity to understand and navigate intricate indoor surroundings. This research advances the creation of more adaptive and intelligent robotic systems that can surmount the constraints of current scene detection models [16].

Indoor scene identification holds considerable importance not just in robotics but also in healthcare [26–28], industrial automation [29, 30], and virtual worlds [31–33]. Precise indoor scene identification improves robots' capacity to execute activities autonomously, engage with their environment, and achieve outcomes more effectively than humans across multiple domains [19, 34–38]. Our research tackles the shortcomings of existing methodologies by introducing an innovative deep CNN model trained on a distinctive dataset. Through rigorous experimentation, we demonstrate the model's efficiency, accuracy, and durability, establishing it as a promising alternative for improving robotic autonomy in practical applications.

Our research presents a comprehensive methodology for interior scene detection in robotic automation, employing a bespoke dataset and deep learning strategies to tackle the complexities of crowded and visually analogous settings. The suggested method demonstrates excellent precision, efficiency, and adaptability, enabling robots to function autonomously in intricate interior environments.

This manuscript is divided into a total of seven distinct sections. The introductory background is presented in Section 1, while prior research contributions and their inadequacies are discussed in Section 2. Section 3 delineates the comprehensive technique, encompassing data collecting, pre-processing, model construction, and parameter configuration. Furthermore, the model training and evaluation have also been included in the same section. Section 4 presents the results obtained by this methodology, while the technical discussion is contained in Section 5. The comprehensive research is encapsulated in section 6-conclusion, accompanied by Section 7, which addresses future directions and recommendations. The entirety of this break-up is illustrated in Figure 1.

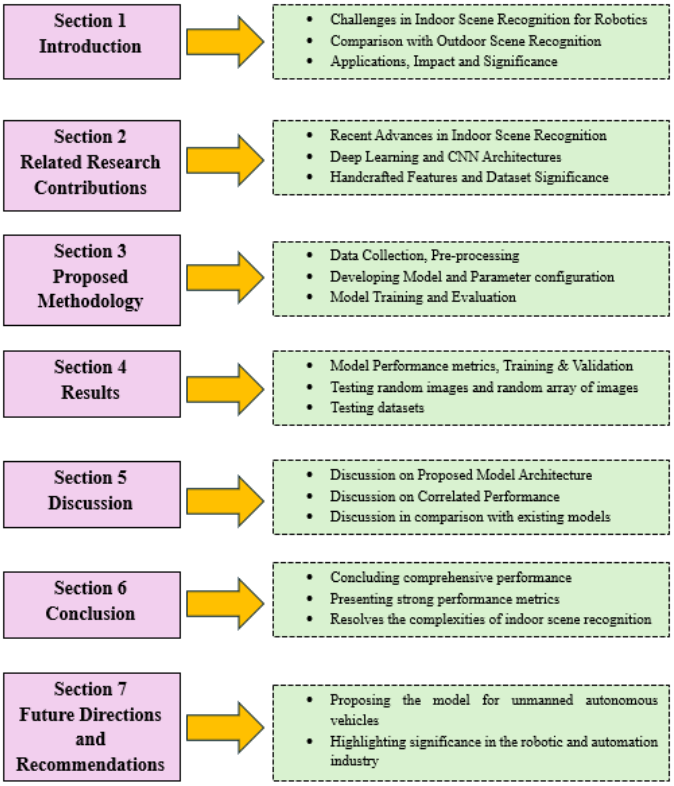

Figure 1. Organization for this research contribution.

2 Related Work

Indoor scene recognition is an important research topic. Recent advancements encompass multisensor models and lightweight deep-learning methodologies. Reference [13, 39–42] introduced a deep learning model that combines 1D Sensor DenseNet with LSTM, attaining an accuracy of over 98%. Afif et al. [3] created a lightweight vision-based detection system tailored for mobile robots, employing weight-trimming techniques and attaining 89% accuracy. Singh et al. [43–48] presented a robust CNN-based methodology for mobile robots, incorporating neuro-fuzzy inference to address uncertainties, achieve an accuracy of 94% and processing 3.1 frames per second.

Deep learning persists in propelling progress [49–52], with CNN architectures such as ResNet [53–55] and attention mechanisms [25] demonstrating remarkable outcomes. Zhou et al. [56] incorporated a multi-scale pyramid pooling module into a ResNet architecture, enhancing the accuracy of the SUN Database. Attention techniques, shown by the multi-scale attention fusion network introduced by Cheng et al. [25], improve discriminative feature extraction, resulting in competitive performance on datasets such as MIT Indoor 67 and NYU Depth V2.

Notwithstanding the emergence of deep learning, conventional handcrafted features remain significant [44]. Specialized datasets such as ADE20K, which include pixel-level annotations for 150 scene categories, are crucial for enhancing semantic segmentation identification tasks. Research persists in investigating innovative techniques for indoor scene recognition. Zhao et al. [22] utilized Building Information Modelling (BIM) for object recognition in interior construction automation, attaining excellent accuracy in real-time item identification. Zhou et al. [27] proposed a Bayesian object relation model to improve contextual object recognition, surpassing alternative methods in indoor scene detection. Miao et al. [55] presented the transfer of object knowledge for scene recognition, with an accuracy of 81.69% on MIT-67. Glavan et al. [5] created InstaIndoor, a multi-modal deep learning methodology, attaining an accuracy of 92.2%. Various research has employed diverse methods, such as Heikel et al.'s [4] integration of object detection with TF-IDF for scene identification, attaining an accuracy of 80.3% on MIT-67. Rafique et al. [21] employed maximum entropy superpixel segmentation and deep belief networks for picture recognition, achieving a 73.85% accuracy on the PASCAL VOC 2012 dataset [53–58]. Anbarasu et al. [59] employed augmented SIFT-ScSPM descriptors for indoor scene detection, attaining an accuracy of 92.2% for micro aerial vehicle navigation. Despite substantial advancements in indoor scene identification algorithms, obstacles persist, especially

Table 1. Studies Related to Indoor Scene Recognition and Understanding.

Reference	Approach	Dataset	Testing Accuracy
[60]	End-to-End CNN	ScanNet v2	88.00%
[56]	Deep Residual Network (ResNet)	SUN Database	75%
[61]	Multi-Scale Attention Fusion Network	MIT Indoor 67 Dataset	80%
[46]	AlexNet, VGGNet, and ResNet	MIT indoor 67 Dataset	94.42%
[62]	VGG-16 Model	MIT Indoor 67 Dataset	83%
[20]	Spatial Pyramid Pooling (SPP) + Deep Learning	SUN Database	88%

in achieving high accuracy across varied datasets. Subsequent research must resolve challenges such as occlusion, dynamic surroundings, and perspective fluctuation to augment the resilience of indoor scene recognition systems. A few recent studies related to indoor scene recognition and understanding are shown in Table 1.

3 Proposed Method

This proposed methodology is based on several stages, as shown in Figure 2, in the form of a research approach. Our research approach has five steps: custom dataset collection, data pre-processing, building model and parameters setting, training data, and evaluation.

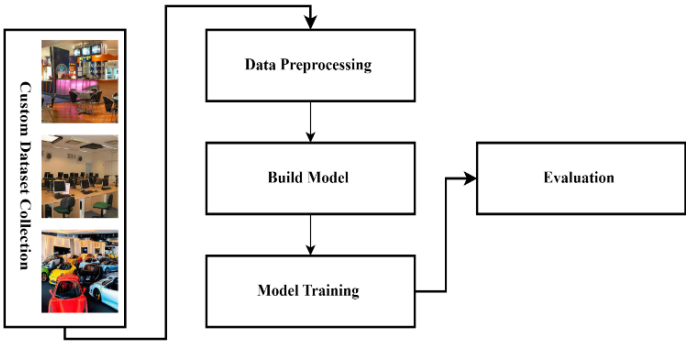

Figure 2. Research Framework.

3.1 Data Collection

A custom dataset has been created for data collection for this research. The images for the custom dataset were downloaded from various web sources and standardized through a series of pre-processing steps. This custom dataset consists of seven indoor scene classes, including the Auditorium, Bar, Bedroom, Car Showroom, Computer Lab, Gym, and Research lab, as shown in Figure 3 on the next page. The custom-created dataset consists of seven classes, and overall, seven thousand images in the dataset, one thousand images of each class, have been collected through different websites. The focus is to create a custom dataset on seven different indoor scenes to train the computer vision's deep learning model to recognize similar objects of various scenes (Computer lab & Research Lab) clearly and cleanly. See details of the custom-created dataset in Table 2.

3.2 Data Pre-processing

In this research, we created a custom dataset, pre-processed it, and successively used it to train a deep convolutional neural network (CNN) sequential model. This custom-created dataset encompasses seven distinct indoor scene classes. In this pre-processing stage, all images have been renamed by executing the Python script on each dataset class and specifying the path to the directory containing images in bulk. Using TensorFlow, the dataset was loaded. Furthermore, image resizing and rescaling were performed. All images were resized to a consistent 256x256 pixel format and organized into batches size of 16 for efficient processing, ensuring uniformity and numerical stability by standardizing pixel values between 0 and 1. Subsequently, the dataset was divided into training (80%), validation (10%), and testing (10%) subsets. Data augmentation techniques, including random horizontal and vertical flips and random rotations of up to 0.2 radians, were applied to improve the model's ability to recognize a broader range of distinct indoor scene classes. Caching and prefetching were employed to expedite data loading and pre-processing during training. The data pre-processing steps included image resizing, rescaling, and data augmentation, which were vital in optimizing the performance of the CNN sequential model for indoor scene recognition and understanding. Sample images are shown in Figure 3 after successfully undergoing the processing steps.

Table 2. Custom created dataset.

Classes	Number	Repository
Auditorium	1000	Custom Created Dataset
Bar	1000	
Bedroom	1000	
Car Showroom	1000	
Computer Lab	1000	
Gym	1000	
Research lab	1000	
Total	7000	

3.3 Build Model and Parameters Setting

The proposed model for this research is carefully designed and sets parameters to work in indoor scene recognition on our custom-created dataset. It adopts a deep Convolutional Neural Network (CNN) with a sequential architecture for indoor scene recognition. This architecture comprises a sequential stack of layers, including convolutional layers with various filter sizes, batch normalization layers for data normalization, max-pooling layers for spatial downsampling, and dense (fully connected) layers for learning complex patterns. Most layers in the model employ the 'relu' activation function, while the output layer uses 'softmax' to predict

(A) (B) (C) (D)

(E) (F) (G)

Figure 3. Images data illustrating various settings: (A) Car Showroom, (B) Bedroom, (C) Computer Lab, (D) Bar, (E) Gym, (F) Auditorium, and (G) Research Lab.

class probabilities. The model is configured with a specified input shape and is tailored to process images with a defined batch size, image size, and multiple channels. The input shape is defined as (BATCH_SIZE, IMAGE_SIZE, IMAGE_SIZE, CHANNELS), where BATCH_SIZE represents the batch size of input images, IMAGE_SIZE signifies the dimensions of input images, and CHANNELS denotes the number of channels in the input data.

The model commences with a "resize and rescale" layer as an initial pre-processing step for input images. Subsequently, multiple convolutional layers are sequentially stacked, each employing various filter sizes and distinct activation functions, 'relu' and 'softmax.' Batch normalization is applied after each convolutional layer to enhance training stability and optimize model performance. Max pooling with a pool size of (2,2) follows each convolutional layer to downsample the feature maps. Additionally, the architecture integrates dense layers with 'relu' activation for further feature extraction and culminates in a final dense layer employing 'softmax' activation for the output of predicted class probabilities. The model's input shape, "input_shape," aligns with the defined (BATCH_SIZE, IMAGE_SIZE, IMAGE_SIZE, CHANNELS) format to seamlessly accommodate the input data.

Additionally, these tailored architectural components work in collaboration to extract essential information from input images progressively and, through meticulous parameter settings, enable the model to recognize distinct indoor scenes precisely. This model structure has been thoughtfully customized to a custom dataset. It aligns with the unique demands of indoor scene recognition within robotic automation, ensuring the system comprehends its environment and responds adeptly to diverse indoor scenarios. The model architecture diagram is shown in Figure 4.

3.4 Model Training

In our model, we've specified a resolution of 256x256 pixels. We utilize Python libraries like TensorFlow, Keras, and Matplotlib for training. The IMAGE_SIZE is set to 256x256, and the BATCH_SIZE is 32, meaning the model processes 32 images simultaneously. CHANNELS are set to 3, indicating that the dataset contains colourful photos. We train the model for 50 EPOCHS, enhancing accuracy confidence and minimizing errors. We use the Adam optimizer and employ the Sparse Categorical Crossentropy loss function for multilevel classification when classes are mutually exclusive. The model was trained using a Kaggle notebook with a P100 GPU and 29 GB of RAM. The model's performance improves with each epoch. It achieves 51% training accuracy in the first epoch, increasing to 66% in the second. As the number of epochs increases, the model's understanding of indoor

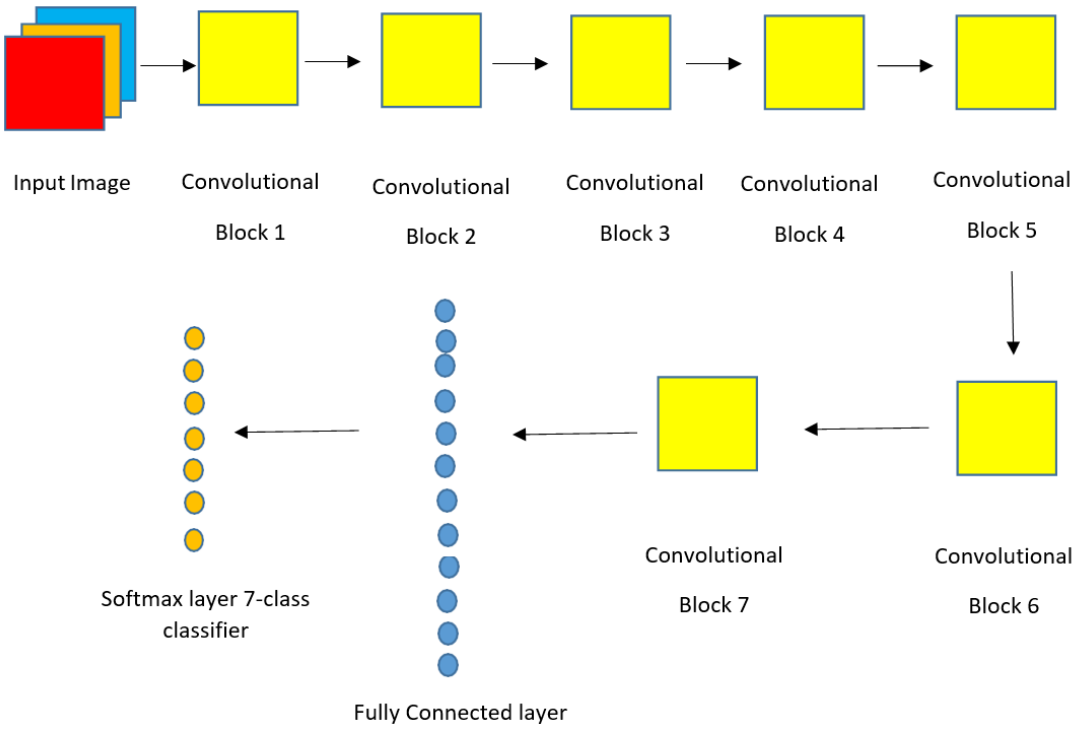

Figure 4. Model Architecture Diagram.

scenes improves. By the fifth epoch, training accuracy reached 86%. This progress demonstrates the model's learning rate enhancement through repeated training sessions.

Furthermore, the error rate or loss consistently decreases during training. By the tenth training iteration, training accuracy reaches about 94%. Subsequently, from the eleventh to the fifteenth iteration, the training accuracy gradually increases by about 2%. At the end of the fifteenth epoch, accuracy is approximately 96%. However, the learning curve exhibits fluctuations starting from the sixteenth epoch, and the accuracy decreases from 98% to 96%. This decline is attributed to the inherent challenges in classifying indoor scenes with similar objects, such as chairs in auditoriums, bars, or computer labs. These ambiguities make distinguishing between such indoor scenes challenging for the model.

At the culmination of the fifty epoch, training accuracy reaches an impressive 99%. This model is trained on a Kaggle notebook with a disk size of up to 73.1GB and a maximum RAM of 13GB. Additionally, it utilizes two T4 GPUs, each with a maximum GPU memory of 14.8GB. Ultimately, the model is trained on a custom dataset and attains a remarkable training accuracy of up to 99%, with a testing accuracy of 89.73%. This achievement represents a significant contribution to indoor scene recognition and understanding.

3.5 Evaluation

In this experiment, the model evaluation becomes a vital measure of how efficiently our model works on our custom-created dataset for training and testing. Next, the model is trained, and the last step is to evaluate it. We evaluated the model performance using measuring metrics, such as accuracy, precision, recall, and F1 score [53]. Formulas (1), (2), (3), and (4) are used to express each metric.

$$\text{Accuracy score} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4}$$

The performance evaluation of the indoor scene recognition model involves several key metrics. True Positives (TP) represent the number of correctly identified instances where the model correctly recognizes an indoor scene category. False Positives (FP) correspond to cases where the model incorrectly identifies an indoor scene category that does not match the actual scene. True Negatives (TN) indicate the count of accurately identified non-matching indoor

scenes. False Negatives (FN) refer to instances where the model fails to recognize a genuine indoor scene, misclassifying it as something else. These metrics are essential for assessing the model's effectiveness in identifying indoor scenes within the robotic automation context.

4 Result

The proposed deep CNN sequential model achieved high training testing accuracy on the custom-created dataset; testing accuracy represents the percentage of correctly predicted samples from a separate. Training accuracy represents the percentage of correctly predicted samples from the training data, indicating the model's ability to classify samples from the dataset accurately. The training accuracy, reaching up to 99%, and testing accuracy, which is 89.73%, 90.11% precision, 89.73% recall, and F1-score of 89.79%, are much better than comparing previous state-of-the-art works, as shown in Table 4, studies conducted on indoor scene recognition and understanding, models and datasets used and trained by other researchers. Thus, our proposed model works well in recognizing and understanding complex indoor scenes on our custom-created datasets. More details are shown in Table 3.

Table 3. Performance metrics for model evaluation on the custom dataset.

Algorithm	Training Accuracy	Testing accuracy	Precision	Recall	F1 Score
CNN-Seq	99%	89.73%	90.11%	89.73%	89.79%

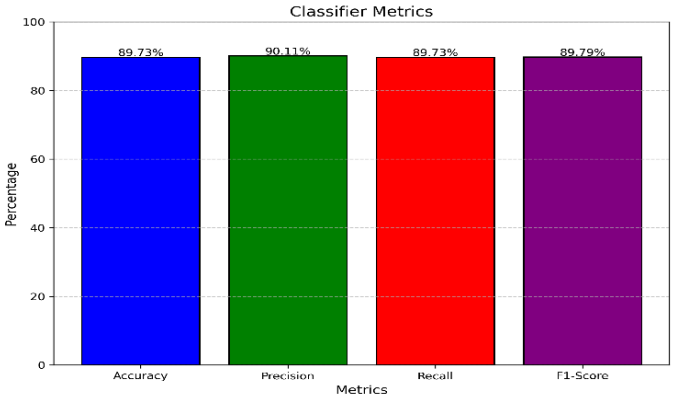

Figure 5. Model Performance Metrics.

The model's performance is shown in Table 3. The testing accuracy reached 89.73%, training accuracy reached 99%, 90.11% precision, 89.11% recall, and an F1-score of 89.79%. These results collectively reflect the strength and reliability of our model on a custom dataset, precisely indoor scene recognition and understanding in robotic automation.

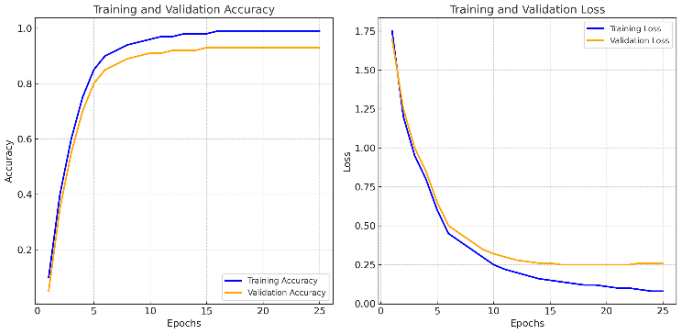

Figure 6. Model Performance MetricsModel Performance Metrics.

Figure 5 shows the model performance metrics developed across different metrics. The bar plot illustrates the performance metrics, including accuracy, precision, recall, and F1 score. Figure 6(a) graph demonstrates the relationship between training accuracy and validation accuracy, showing that training accuracy increases exponentially with the number of epochs, reaching 99% by the 50th epoch. Similarly, validation accuracy initially increases rapidly in the first five epochs but fluctuates between 80% and 94% after that; due to similarities in the dataset images, final validation accuracy is recorded as 92%, as shown in Figure 6(a) graph. Figure 6(b) graph illustrates the relationship between training loss and validation loss during the model training process. Training loss, representing the error between predicted and actual values, decreases exponentially as the number of epochs increases, reaching 0.15 from 1.34 by the tenth epoch and 0.01 by the 50th epoch. Similarly, validation loss initially decreases rapidly from 2.94 to 0.57 in the first ten epochs but then fluctuates between 0.47 and 0.67 due to similarities in the dataset images. Ultimately, the validation loss is recorded as 0.39 at the end of the training process.

Figure 7 demonstrates the model's accurate predictions for all three scenes. The first scene, a Computer Lab, aligns with the model's prediction. Similarly, the second scene, a Research Lab, is correctly identified as such by the model. The third scene, another Computer Lab, is also accurately classified by the model.

Figure 8 demonstrates the model's accurate predictions for all three scenes. The first scene is identified as a Computer Lab, as is the model's prediction. The second scene is labelled a Bar, matching the model's prediction. The third scene is recognized as a Research Lab, which aligns with the model's prediction.

Figure 7. Testing results in a random array of images taken from the test dataset.

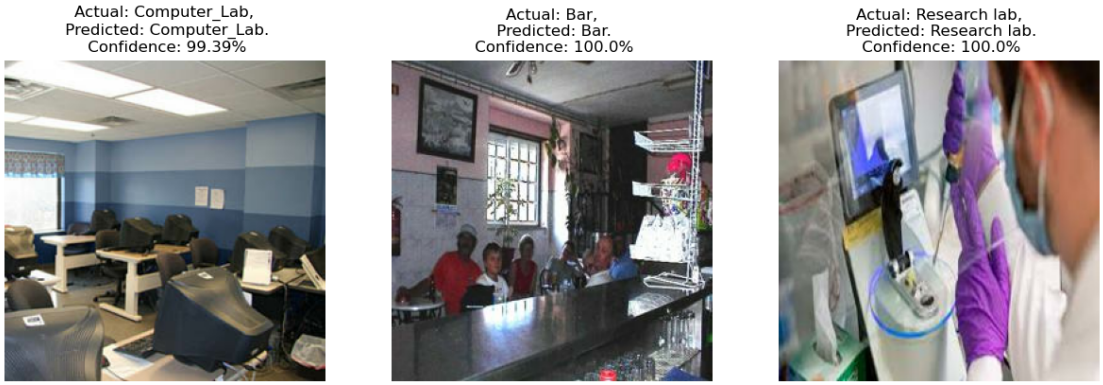

Figure 8. Testing result from random images taken from the test dataset.

The proposed model in this research achieves a high training accuracy of 99% and testing accuracy of 89.73% on our custom dataset containing seven different indoor scene classes, outperforming other existing models. Yue et al. [60] achieved a testing accuracy of 88% on the ScanNet v2 Database using the End-to-End CNN. Zhou et al. [56] achieved a testing accuracy of approximately 75% on the SUN Database using the ResNet model. Chen et al. [61] reported an 80% testing accuracy on the MIT Indoor 67 dataset using a multi-scale attention fusion network. Afif et al. [46] achieved a high testing accuracy of 94.42% on the MIT 67 indoor Dataset using AlexNet, VGGNet, and ResNet. Kim et al. [62] reported an 83% testing accuracy on the MIT Indoor 67 dataset using the VGG-16 model. Liu et al. [20] achieved a high testing accuracy of 88% on the SUN Database using the Spatial Pyramid Pooling (SPP) and deep learning approach. So, the suggested model performs better than most current models, demonstrating its potency in indoor scene detection.

Figure 9 illustrates the testing mechanism in which random images from the separate testing dataset, comprising 10% of the total dataset, are selected. These images are then tested against the trained dataset to predict the scene.

Table 4. Proposed model results Comparison with previous state-or-art works.

References	Approach	Dataset	Testing Accuracy
[60]	End-to-End CNN	ScanNet v2	88.00%
[56]	Deep Residual Network (ResNet)	SUN Database	75%
[61]	Multi-Scale Attention Fusion Network	MIT Indoor 67 Dataset	80%
[46]	AlexNet, VGGNet, and ResNet	MIT 67 indoor Dataset	94.42%
[62]	VGG-16 Model	MIT Indoor 67 Dataset	83%
[20]	Spatial Pyramid Pooling (SPP) + Deep Learning	SUN Database	88%
Our model	**CNN-Seq Model**	**Custom created dataset**	**89.73%**

Figure 9. Testing result taken from the test dataset.

5 Discussion

This research introduces a comprehensive method for indoor scene detection with a bespoke dataset and a deep convolutional neural network sequential model. The dataset comprises seven interior scene categories, with 80% allocated for training and 10% each for validation and testing. The proposed model attains a training accuracy of 99% and a testing accuracy of 89.73%. The design employs a deep convolutional neural network using sequential layers comprising dense layers, max-pooling, batch normalization, and convolutional layers with diverse filter sizes. The 'softmax' activation function is utilized at the output layer, whilst 'relu' is employed throughout the model. This architecture facilitates effective feature extraction and classification, accommodating images with designated batch sizes, dimensions, and channels. The model exhibited robust performance metrics: 90.11% precision, 89.73% recall, and an F1-score of 89.79%.

Our method surpasses other models [54, 60, 63]. Yue et al. [60] attained 88% testing accuracy on the ScanNet v2 Database, whilst Zhou et al. [56] documented 75% accuracy utilizing ResNet on the SUN Database. Chen et al. [61] attained an accuracy of 80% on the MIT Indoor 67 dataset. Conversely, our model achieved a superior accuracy of 89.73%. Furthermore, our model surpasses the performance of Liu et al. [20], who attained 88% on the SUN Database, and S. Jeong Kim et al. [62], who secured 83% on MIT Indoor 67 utilizing VGG-16. Subsequent comparisons indicate that our model exceeds the results of Miao et al. [55] (72.5% on MIT-67), Glavan et al. [5](79.8% on SUN-397), and Heikel et al. [4] (71.8% on MIT-67). Rafique et al. [21] obtained 77.4% on SUN-397, but Anbarasu et al. [59] secured 72.4% on the UC Merced dataset. The model's enhanced accuracy, relative to these benchmarks, highlights its applicability in real-world indoor scene detection.

6 Conclusion

A custom dataset including seven distinct interior scene classes and images demonstrated strong performance for the proposed deep learning model in indoor scene detection and recognition. The model outperformed the earlier models employed by other researchers and AI experts, achieving testing accuracy of 89.73%, training accuracy of up to 99%, 90.11% precision, 89.73% recall, and an F1-score of 89.79%. The model was successfully implemented using the deep learning CNN sequential layers model with unique parameter settings. As indoor scenes are designed for the benefit of humans, artificial intelligence-based systems or models are widely utilized in everyday life, including the significant fields of machine learning and deep learning. The main objective is to further enhance the existing model and custom dataset by adding different types of classes and images. It is essential to develop new models and datasets to address the challenges of indoor scene recognition and understanding, which are more complex than outdoor scene recognition. This research highlights the significance of gaining knowledge about indoor scenes and their objects, composition, configuration, and classification to support humans or robots and any actions they perform within that space. Numerous innovative and unique solutions have been explored in this field.

7 Future Directions and Recommendations

This research will benefit other researchers researching this excellent and challenging topic. Such models can be trained with a massive number of classes in custom datasets to make the machines more intelligent to behave and analyze any scene anywhere and in any environment efficiently and effectively; just like humans, they could be able to make rational and sound decisions by analyzing and understanding all types of indoor scenes. This model is limited to recognizing and understanding specific indoor scenes. However, in the future, it can be made from specific to general if this model can be trained on a massive number of classes in this custom dataset and on other datasets covering a wide range of all real-life indoor scenes.

Conflicts of Interest

The authors declare no conflicts of interest.

Funding

This work was jointly supported by the Data and Intelligence Laboratory (D&Intel Lab), School of Computer Science and Engineering, Southeast University, China and the Robotics Control lab under the Interdisciplinary Research Centre for Aviation and Space Exploration (IRC-ASE), King Fahd University of Petroleum and Minerals (KFUPM), Kingdom of Saudi Arabia.

References

[1] Macrorie, R., Marvin, S., & While, A. (2021). Robotics and automation in the city: a research agenda. *Urban Geography*, 42(2), 197-217. [CrossRef]

[2] Kolpashchikov, D., Gerget, O., & Meshcheryakov, R. (2022). Robotics in healthcare. Handbook of Artificial Intelligence in Healthcare: *Vol 2: Practicalities and Prospects*, 281-306. [CrossRef]

[3] Afif, M., Ayachi, R., Said, Y., & Atri, M. (2022). An evaluation of EfficientDet for object detection used for indoor robots assistance navigation. *Journal of Real-Time Image Processing*, 19(3), 651-661. [CrossRef]

[4] Heikel, E., & Espinosa-Leal, L. (2022). Indoor scene recognition via object detection and TF-IDF. Journal of Imaging, 8(8), 209. [CrossRef]

[5] Glavan, A., & Talavera, E. (2022). InstaIndoor and multi-modal deep learning for indoor scene recognition. *Neural Computing and Applications*, 34(9), 6861-6877. [CrossRef]

[6] Fang, W., Chen, L., Zhang, T., Chen, C., Teng, Z., & Wang, L. (2023). Head-mounted display augmented reality in manufacturing: A systematic review. *Robotics and Computer-Integrated Manufacturing*, 83, 102567. [CrossRef]

[7] Khan, S. H., Hayat, M., Bennamoun, M., Togneri, R., & Sohel, F. A. (2016). A discriminative representation of convolutional features for indoor scene recognition. *IEEE Transactions on Image Processing*, 25(7), 3372-3383. [CrossRef]

[8] Khan, S. D., & Othman, K. M. (2024). Indoor Scene Classification through Dual-Stream Deep Learning: A Framework for Improved Scene Understanding in Robotics. *Computers*, 13(5), 121. [CrossRef]

[9] Li, X. (2024, April). Intelligent Inspection Robot Scene Recognition under Convolutional Neural Network. In *2024 IEEE 13th International Conference on Communication Systems and Network Technologies* (*CSNT*) (pp. 519-524). IEEE. [CrossRef]

[10] Santos, D., Lopez-Lopez, E., Pardo, X. M., Iglesias, R., Barro, S., & Fdez-Vidal, X. R. (2019). Robust and fast scene recognition in robotics through the automatic identification of meaningful images. *Sensors*, 19(18), 4024. [CrossRef]

[11] Dahri, F. H., Dahri, N. A., & Soomro, M. A. (2023). Image caption generator using convolutional recurrent neural network feature fusion. *Journal of Xi'an Shiyou University, Natural Science Edition*, 9, 1088-1095.

[12] Sharma, V., Nagpal, N., Shandilya, A., Dureja, A., & Dureja, A. (2022, December). A Practical Approach to detect Indoor and Outdoor Scene Recognition. In *Proceedings of the 4th International Conference on Information Management & Machine Intelligence* (pp. 1-10). [CrossRef]

[13] Zhu, Y., Luo, H., Zhao, F., & Chen, R. (2020). Indoor/outdoor switching detection using multisensor DenseNet and LSTM. *IEEE Internet of Things Journal*, 8(3), 1544-1556. [CrossRef]

[14] Kuriakose, B., Shrestha, R., & Sandnes, F. E. (2021, October). SceneRecog: a deep learning scene recognition model for assisting blind and visually impaired navigate using smartphones. In *2021 IEEE International Conference on Systems, Man, and Cybernetics* (*SMC*) (pp. 2464-2470). IEEE. [CrossRef]

[15] Alqobali, R., Alshmrani, M., Alnasser, R., Rashidi, A., Alhmiedat, T., & Alia, O. M. D. (2023). A survey on robot semantic navigation systems for indoor environments. *Applied Sciences*, 14(1), 89. [CrossRef]

[16] Wijayathunga, L., Rassau, A., & Chai, D. (2023). Challenges and solutions for autonomous ground robot scene understanding and navigation in unstructured outdoor environments: A review. *Applied Sciences*, 13(17), 9877. [CrossRef]

[17] Daou, A., Pothin, J. B., Honeine, P., & Bensrhair, A. (2023). Indoor scene recognition mechanism based on direction-driven convolutional neural networks. *Sensors*, 23(12), 5672. [CrossRef]

[18] Kumar, N., Singh, H., Varshney, M. T., Malik, M. V., & Kumar, V. (2022). Indoor and Outdoor Scene Recognition. *Grenze International Journal of Engineering & Technology* (*GIJET*), 8(2).

[19] Georgiou, A., Masters, P., Johnson, S., & Feetham, L. (2022). UAV-assisted real-time evidence detection in outdoor crime scene investigations. *Journal of forensic sciences*, 67(3), 1221-1232. [CrossRef]

[20] Liu, S., & Tian, G. (2019). An indoor scene classification method for service robot Based on CNN feature. *Journal of Robotics*, 2019(1), 8591035. [CrossRef]

[21] Rafique, A. A., Gochoo, M., Jalal, A., & Kim, K. (2023). Maximum entropy scaled super pixels segmentation for multi-object detection and scene recognition via deep belief network. *Multimedia Tools and Applications*, 82(9), 13401-13430. [CrossRef]

[22] Zhao, X., & Cheah, C. C. (2023). BIM-based indoor mobile robot initialization for construction automation using object detection. *Automation in Construction*, 146, 104647. [CrossRef]

[23] Wang, H., & Li, M. (2024). A new era of indoor scene reconstruction: A survey. *IEEE Access*, 12, 110160-110192. [CrossRef]

[24] Choe, S., Seong, H., & Kim, E. (2021). Indoor place category recognition for a cleaning robot by fusing a probabilistic approach and deep learning. *IEEE Transactions on Cybernetics*, 52(8), 7265-7276. [CrossRef]

[25] Cheng, C., Koschan, A., Chen, C. H., Page, D. L., & Abidi, M. A. (2011). Outdoor scene image segmentation based on background recognition and perceptual organization. *IEEE transactions on image processing*, 21(3), 1007-1019. [CrossRef]

[26] Gupta, S., Arbelaez, P., & Malik, J. (2013). Perceptual organization and recognition of indoor scenes from RGB-D images. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 564-571).

[27] Zhou, Z., Li, L., Fürsterling, A., Durocher, H. J., Mouridsen, J., & Zhang, X. (2022). Learning-based object detection and localization for a mobile robot manipulator in SME production. *Robotics and Computer-Integrated Manufacturing*, 73, 102229. [CrossRef]

[28] Samani, E. U., Yang, X., & Banerjee, A. G. (2021). Visual object recognition in indoor environments using topologically persistent features. *IEEE Robotics and Automation Letters*, 6(4), 7509-7516. [CrossRef]

[29] Silvera-Tawil, D. (2024). Robotics in Healthcare: A Survey. *SN Computer Science*, 5(1), 189. [CrossRef]

[30] Liu, M., Chen, M., Wu, Z., Zhong, B., & Deng, W. (2024). Implementation of Intelligent Indoor Service Robot Based on ROS and Deep Learning. *Machines*, 12(4), 256. [CrossRef]

[31] Strader, J., Hughes, N., Chen, W., Speranzon, A., & Carlone, L. (2024). Indoor and outdoor 3d scene graph generation via language-enabled spatial ontologies. *IEEE Robotics and Automation Letters*, 9(6), 4886-4893. [CrossRef]

[32] Liu, Z., Wang, J., Li, J., Liu, P., & Ren, K. (2023). A novel multiple targets detection method for service robots in the indoor complex scenes. *Intelligent Service Robotics*, 16(4), 453-469. [CrossRef]

[33] Von Itzstein, G. S., Billinghurst, M., Smith, R. T., & Thomas, B. H. (2024). Augmented reality entertainment: Taking gaming out of the box. In *Encyclopedia of computer graphics and games* (pp. 162-170). Cham: Springer International Publishing. [CrossRef]

[34] Feng, J., Sun, J., & Yao, Y. (2023, April). Design of Intelligent Service Robot for Military Recuperation. In *2023 IEEE International Conference on Control, Electronics and Computer Technology (ICCECT)* (pp. 131-137). IEEE. [CrossRef]

[35] Heckelman, L. N., Soher, B. J., Spritzer, C. E., Lewis, B. D., & DeFrate, L. E. (2022). Design and validation of a semi-automatic bone segmentation algorithm from MRI to improve research efficiency. *Scientific Reports*, 12(1), 7825. [CrossRef]

[36] Borrego, A., Latorre, J., Alcañiz, M., & Llorens, R. (2018). Comparison of Oculus Rift and HTC Vive: feasibility for virtual reality-based exploration, navigation, exergaming, and rehabilitation. *Games for health journal*, 7(3), 151-156. [CrossRef]

[37] Maruhn, P. (2021). VR Pedestrian Simulator Studies at Home: Comparing Google Cardboards to Simulators in the Lab and Reality. *Frontiers in Virtual Reality*, 2, 746971. [CrossRef]

[38] Lee, C. D. (2021). A Review of Virtual and Augmented Reality Concepts, Technologies and Application. Journal of Computing and Natural Science, 1(4), 139-144.

[39] Qi, J., Ma, L., Cui, Z., & Yu, Y. (2024). Computer vision-based hand gesture recognition for human-robot interaction: a review. *Complex & Intelligent Systems*, 10(1), 1581-1606. [CrossRef]

[40] Bhola, G., & Vishwakarma, D. K. (2024). A review of vision-based indoor HAR: state-of-the-art, challenges, and future prospects. *Multimedia Tools and Applications*, 83(1), 1965-2005. [CrossRef]

[41] Emek Soylu, B., Guzel, M. S., Bostanci, G. E., Ekinci, F., Asuroglu, T., & Acici, K. (2023). Deep-learning-based approaches for semantic segmentation of natural scene images: A review. Electronics, 12(12), 2730. [CrossRef]

[42] Ismail, A. S., Seifelnasr, M. M., & Guo, H. (2018, April). Understanding indoor scene: Spatial layout estimation, scene classification, and object detection. In *Proceedings of the 3rd International Conference on Multimedia Systems and Signal Processing* (pp. 64-70). [CrossRef]

[43] Sitaula, C., Xiang, Y., Zhang, Y., Lu, X., & Aryal, S. (2019). Indoor image representation by high-level semantic features. *IEEE Access*, 7, 84967-84979. [CrossRef]

[44] Susan, S., & Tuteja, M. (2024). Feature Engineering Versus Deep Learning for Scene Recognition: A Brief Survey. *International Journal of Image and Graphics*, 2550054. [CrossRef]

[45] Guo, J., Chen, H., Liu, B., & Xu, F. (2023). A system and method for person identification and positioning incorporating object edge detection and scale-invariant feature transformation. *Measurement*, 223, 113759. [CrossRef]

[46] Afif, M., Ayachi, R., Said, Y., & Atri, M. (2020). Deep learning based application for indoor scene recognition. *Neural Processing Letters*, 51, 2827-2837. [CrossRef]

[47] Surendran, R., Chihi, I., Anitha, J., & Hemanth, D. J. (2023). Indoor Scene Recognition: An Attention-Based Approach Using Feature Selection-Based Transfer Learning and Deep Liquid State Machine. *Algorithms*, 16(9), 430. [CrossRef]

[48] Singh, A., Pandey, P., Puig, D., Nandi, G. C., & Abdel-Nasser, M. (2022). Reliable Scene Recognition Approach for Mobile Robots with Limited Resources Based on Deep Learning and Neuro-Fuzzy Inference. *Traitement du Signal*, 39(4), 1255.

[49] Quattoni, A., & Torralba, A. (2009, June). Recognizing indoor scenes. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 413-420). IEEE. [CrossRef]

[50] Bose, D., Hebbar, R., Somandepalli, K., Zhang, H., Cui, Y., Cole-McLaughlin, K., ... & Narayanan, S. (2023). Movieclip: Visual scene recognition in movies. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 2083-2092).

[51] Velikov, K. (2023). *Enhancing Semantic Segmentation for Indoor Environments: Integrating Depth Information into Neural Networks* (Bachelor's thesis, University of Twente).

[52] Piekenbrinck, J., Hermans, A., Vaskevicius, N., Linder, T., & Leibe, B. (2024). RGB-D Cube R-CNN: 3D Object Detection with Selective Modality Dropout. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 1997-2006).

[53] Naidu, G., Zuva, T., & Sibanda, E. M. (2023, April). A review of evaluation metrics in machine learning algorithms. In *Computer Science On-line Conference* (pp. 15-25). Cham: Springer International Publishing. [CrossRef]

[54] Javed, M., Zhang, Z., Dahri, F. H., & Laghari, A. A. (2024). Real-time deepfake video detection using eye movement analysis with a hybrid deep learning approach. *Electronics*, 13(15), 2947. [CrossRef]

[55] Miao, B., Zhou, L., Mian, A. S., Lam, T. L., & Xu, Y. (2021, September). Object-to-scene: Learning to transfer object knowledge to indoor scene recognition. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 2069-2075). IEEE. [CrossRef]

[56] Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2017). Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6), 1452-1464. [CrossRef]

[57] Zhou, Z., Zhang, J., Gong, C., & Wu, W. (2023). Automatic tunnel lining crack detection via deep learning with generative adversarial network-based data augmentation. *Underground Space*, 9, 140-154. [CrossRef]

[58] Xie, L., Lee, F., Liu, L., Kotani, K., & Chen, Q. (2020). Scene recognition: A comprehensive survey. *Pattern Recognition*, 102, 107205. [CrossRef]

[59] Anbarasu, B., & Anitha, G. (2018). Indoor scene recognition for micro aerial vehicles navigation using enhanced-GIST descriptors. *Defence Science Journal*, 68(2), 129.

[60] Yue, H., Lehtola, V., Wu, H., Vosselman, G., Li, J., & Liu, C. (2024). Recognition of Indoor Scenes using 3D Scene Graphs. IEEE Transactions on Geoscience and Remote Sensing. [CrossRef]

[61] Chen, H., Li, Y., & Su, D. (2019). Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for RGB-D salient object detection. *Pattern Recognition*, 86, 376-385. [CrossRef]

[62] Kim, S. J., & Shin, D. H. (2017, January). The effects of ambient scent on hedonic experience on online shopping. In *Proceedings of the 11th international conference on ubiquitous information management and communication* (pp. 1-5). [CrossRef]

[63] Dahri, F. H. (2022). Automatic Face Mask Detection and Recognition Using Deep Learning. *ScienceOpen Preprints*, 13(11), 433-447. [CrossRef]

Fida Hussain Dahri received his Bachelor's degree in Information Technology from Quaid-e-Awam University of Engineering, Sciences & Technology Nawabshah, Pakistan, in 2020, securing 3rd position in his faculty. He completed his Master's degree in Information Technology from the same university in 2023, graduating in the top 2% of his class. He is pursuing a second Master's in Computer Science and Technology at Southeast University, Nanjing, China, where he is also a Research Assistant at the Data and Intelligence Laboratory (D&Intel Lab). His research focuses on computer vision, machine learning, deep learning, and image processing. He has worked in artificial intelligence and computer vision and continues contributing to advancements in these fields. (Email: 223227084@seu.edu.cn)

Ghulam E Mustafa Abro earned his B.S. in Electronic Engineering with honours from Hamdard University, Pakistan, in 2016, followed by an M.S. in Control and Automation from Sir Syed University in 2019 and a PhD in Electrical and Electronic Engineering from Universiti Teknologi PETRONAS, Malaysia, in 2023. He is a postdoctoral fellow at King Fahd University of Petroleum and Minerals (KFUPM) in Saudi Arabia, working at the Interdisciplinary Research Centre for Aviation and Space Exploration. Dr. Abro has nearly a decade of involvement with IEEE, serving in various roles, including conference chair and reviewer for SCI-indexed journals. His diverse research interests span control of underactuated systems, autonomous navigation, robotics, swarm technology, and multi-agent systems. Before KFUPM, he held academic and research roles at Hamdard University, Universiti Teknologi PETRONAS, and defence research institutes in Malaysia. (Email: Ghulam.abro@kfupm.edu.sa)

Nisar Ahmed Dahri received a Ph.D. in information technology from the Quaide-Awam University of Engineering, Science and Technology, Nawabshah, Pakistan. He works as a postdoctoral fellow at Universiti Teknologi Malaysia (UTM), Malaysia, and he is a visiting faculty member at the Department of Education, SBBU University Shaheed Benazirabad. His research area is ICT in education, Mobile learning, MOOCs, eLearning, Technology Adoption Models, and HCI. (Email: ahmeddahri.nisar@utm.my)

Asif Ali Laghari received a B.S. degree in Information Technology from the Quaid-e-Awam University of Engineering Science and Technology Nawabshah, Pakistan, in 2007 and a Master's degree in Information Technology from the QUEST Nawabshah Pakistan in 2014. From 2007 to 2008, he was a Lecturer in the Computer and Information Science Department at the Digital Institute of Information Technology, Pakistan. In 2015, he joined the School of Computer Science & Technology, Harbin Institute of Technology, where he was a Ph.D. student. He is an assistant professor at Sindh Madressatul Islam University, Karachi, Pakistan, and is affiliated with Shenyang Normal University. He has published over 130 technical articles in scientific journals and conference proceedings. His current research interests include Machine Learning, Computer networks, cloud computing, IoT, Fog computing, and multimedia QoE management. (Email: Asiflaghari@synu.edu.cn)

Zain Anwar Ali earned his B.S. in Electronic Engineering from Sir Syed University of Engineering and Technology (SSUET), Karachi, in 2009, followed by an M.S. in Industrial Control and Automation from Hamdard University in 2012 and a Ph.D. in Control Theory and Engineering from Nanjing University of Aeronautics and Astronautics (NUAA) in 2017. He has held academic positions at SSUET and Hamdard University, & conducted PhD research with Nanjing Strong Flight Electronics. He is an assistant professor at the Electronic Engineering Department of Maynooth International Engineering College (MIEC), Maynooth University, Maynooth, Co. Kildare, Ireland. Dr Ali has published over 73 research articles and is a member of various international engineering bodies. The Chinese Ministry twice selected him as a Highly Talented Foreign Expert. He has served as Assistant Editor of SSUET Research Journal and Director of the Continuing Education Program at SSUET, and he participates in research collaborations funded by Pakistan's Higher Education Commission (HEC). (Email: Zainanwar.ali@mu.ie)