algorithms

MDPI

Article

Indoor Scene Recognition: An Attention-Based Approach Using Feature Selection-Based Transfer Learning and Deep Liquid State Machine

Ranjini Surendran [1], Ines Chihi [2], J. Anitha [2] and D. Jude Hemanth [1,*]

[1] Department of ECE, Karunya Institute of Technology and Sciences, Coimbatore 641114, India; ranjinirajeshk@gmail.com
[2] Department of Engineering, Faculty of Science, Technology and Medicine, University of Luxembourg, 1359 Luxembourg, Luxembourg; ines.chihi@uni.lu (I.C.); anithaj@karunya.edu (J.A.)
* Correspondence: judehemanth@karunya.edu

Abstract: Scene understanding is one of the most challenging areas of research in the fields of robotics and computer vision. Recognising indoor scenes is one of the research applications in the category of scene understanding that has gained attention in recent years. Recent developments in deep learning and transfer learning approaches have attracted huge attention in addressing this challenging area. In our work, we have proposed a fine-tuned deep transfer learning approach using DenseNet201 for feature extraction and a deep Liquid State Machine model as the classifier in order to develop a model for recognising and understanding indoor scenes. We have included fuzzy colour stacking techniques, colour-based segmentation, and an adaptive World Cup optimisation algorithm to improve the performance of our deep model. Our proposed model would dedicatedly assist the visually impaired and blind to navigate in the indoor environment and completely integrate into their day-to-day activities. Our proposed work was implemented on the NYU depth dataset and attained an accuracy of 96% for classifying the indoor scenes.

Keywords: deep learning; DenseNet; fuzzy colour stacking; liquid state machine; transfer learning; world cup optimization

Citation: Surendran, R.; Chihi, I.; Anitha, J.; Hemanth, D.J. Indoor Scene Recognition: An Attention-Based Approach Using Feature Selection-Based Transfer Learning and Deep Liquid State Machine. *Algorithms* **2023**, *16*, 430. https://doi.org/10.3390/a16090430

Academic Editors: Arslan Munir and Frank Werner

Received: 6 August 2023
Revised: 28 August 2023
Accepted: 5 September 2023
Published: 8 September 2023

1. Introduction

Understanding a scene's environment is an effective task for humans. We recognise a scene with the knowledge we gain through our short-term or long-term observations and experiences in similar situations. Making a machine recognise a scene as a human does has tremendous applications in the field of robotics. When it comes to machines, recognising a scene by understanding the semantic-level details is a challenging task. Recognising the different scene categories, which include outdoor scenes, indoor scenes, human activity scenes, aerial scenes, etc., and identifying the different objects in the scene while preserving their semantic relations is still a great challenge in many state-of-the-art approaches. The advances in artificial intelligence, with the great success of deep learning, have elevated its performance in many computer-vision applications such as object detection, scene classification [1], face recognition, human-action recognition [2], etc. Scene classification or recognition is one of the most significant and challenging tasks in the area of robotics [3] and computer vision. Unlike the object recognition task, which consists of images having single or multiple occurrences of the same object, scene recognition is difficult since a scene involves the occurrence of multiple dissimilar objects. Indoor scene recognition has attracted great attention in recent years due to its widespread application in numerous areas, including smart navigation systems, elderly monitoring systems, domestic robotics, intelligent surveillance, etc.

A scene could be either dynamic or static. Dynamic scenes include the video scene [4], and static scenes include the still image scenes. Recognising static scenes is more complicated than recognising dynamic scenes. This is because, in order to recognise an object in a video, we need a number of frames, compared to the partial information obtained from still images. Static scene recognition can be classified into two categories: outdoor scene recognition and indoor scene recognition. Outdoor scene recognition involves recognising scenes like a street, forest, beach, etc. that consist of more or less similar objects in a scene. But indoor scene recognition involves scenes like the living room, classroom, bedroom, etc., which consist of different objects with varying dimensions, positions, and occlusions. Indoor environments provide rich decorations, complex occlusions, different textures and scales, and a cluttered background, making it a challenging task in the area of scene recognition.

In our society, we have a large population of people suffering from visual ailments, ranging from mild to severe visual impairments. Indoor scene recognition and understanding can effectively assist the visually impaired, sighted, and elderly in navigating through their indoor environment by analysing their surrounding environment. In our paper, we have proposed a hybrid model using the concept of transfer learning-based deep approaches for understanding indoor scenes that can provide assistance for indoor navigation.

1.1. Challenges and Motivation

Much research has been carried out in dynamic scene recognition, as the semantic information among the different objects in a scene is preserved by the different frames of the video, and the orientation information is not lost. Classification of the still image scene is a challenging task due to the lack of spatial information about multiple objects in the still image. Also, the lack of a sufficient still image dataset is another challenge in this field. Many of the state-of-the-art approaches use techniques like data augmentation and image annotation to improve their performance, which results in a reduced diversity of data and is time-consuming. Although the majority of research works focus on the recognition of outdoor scenes, recent research has focused on indoor scene recognition. Many state-of-the-art approaches that have high performance in outdoor scenes do not have the same performance in indoor scenes. Unlike the outdoor scene, the indoor scene suffers from many challenges, including variation in illumination, size, spatial layout, several occlusions, varying spatial locations, etc.

Indoor scene images consist of several discrete objects with diverse appearances, sizes, shapes, poses, and scales that can be placed at various spatial locations in a multitude of possible layouts. The major challenges in indoor scene recognition are the large-scale variations in spatial scaling and spatial layout of the different objects in the scene. For the above-mentioned challenges, firstly, in different scenes of the same class, the constituent objects have different spatial scales. Secondly, these constituent objects may be located in varying spatial locations and in numerous possible layouts. For example, a fridge can be available in a kitchen at different spatial locations with significant differences in appearance and scale. Some indoor scenes can be characterised by global appearance information, whereas others require local spatial information. For example, a corridor scene can be characterised by a single object (e.g., a large wall), but a kitchen scene can include multiple objects (e.g., a fridge, an oven, a cooking range). Therefore, in order to accommodate all the diverse scene types, both local and global spatial information needs to be considered. This is also a critical condition for indoor scene recognition. Indoor scene environments consist of rich and decorative features in disordered patterns that vary in size, orientation, and layout. The above-mentioned challenges inspired us to work on indoor scene recognition in order to achieve invariance in the scaling and spatial layout of the objects in the indoor scene images.

1.2. Contributions

In this paper, we have proposed a fine-tuned pretrained deep framework model to increase the performance of existing indoor scene recognition tasks. In this work, we have worked on an NYU depth dataset consisting of different classes of indoor scenes. The images are preprocessed using fuzzy colour stacking techniques. These preprocessed images are segmented and then fed to the pretrained DenseNet201 model for feature extraction. A feature selection process using adaptive World Cup optimisation selects the most superior features and feeds them to a deep classifier known as an LSM (liquid state machine).

Some of the major contributions of our work can be summarised as follows:

- Implementation of fuzzy colour stacking for preprocessing improves the foreground quality of the images by filtering the background noise;
- Unlike outdoor scene recognition that utilises global spatial information, indoor scene recognition is possible using the objects in the scene. Therefore, for both local and global features, semantic information is needed. Hence, the segmentation provides an ROI (region of interest) to detect the objects;
- Pretrained DenseNet201 improves feature extraction due to a lack of vanishing gradient problems;
- An attention module to select the best features using the World Cup optimisation algorithm improves the robust nature of the model towards indoor scene recognition;
- Classification using the deep LSM model utilising the winner-take-all layer improves the overall accuracy of the indoor scene classification.

We have structured the paper as follows: In Section 2, some of the existing works and a literature survey of indoor scene recognition are discussed. In Section 3, we have explained the block diagram and workflow of our proposed approach. In Section 4, we have given details on the dataset and some sample indoor scenes. Section 5 shows the experimental results and discusses the performance matrices of our classifier model, and, in Section 6, we compare the performance of our model with existing indoor scene recognition works. Section 7 includes our concluding remarks with the future scope.

2. Related Works

Indoor scene recognition has gained great attention in recent years, and many research projects have been developed in this field. Most methods of scene recognition consist of a three-step approach, which consists of learning the features in different positions and scales of the objects in the scene, pooling or cumulating these learned features in order to obtain the feature description, and, based on the feature representation, learning a suitable classifier. Traditional indoor scene recognition techniques mainly focus on global features of the scene. In these approaches, low-level features like edges, colours, and textures [5] are used for scene classification. For the initial step, the traditional state-of-the-art methods used handcrafted methods like SIFT [6] (scale invariant feature transformation), SURF [7] (speeded up robust features), GIST [8], and HOG [9] (histogram of oriented gradients). Later, mid-level feature extraction using BoVW [10] (Bag of Visual Words) was employed to improve the extracted features. In the next phase, these extracted local descriptors with different scales and locations are aggregated using encoding pooling methods such as FV [11] (Fisher Vector) and VLAD [12] (Vector of Locally Aggregated Descriptors). In the last phase, based on the feature representation, some widely used classifiers such as SVM (Support Vector Machine), KNN (K-nearest neighbour), neural networks, etc. are employed for the scene recognition task. These traditional methods employ the use of handcrafted image features for classification.

Deep learning [13] has become a better solution to improve the performance of many machines in understanding specific computer vision tasks. Unlike conventional machine-learning approaches, where the features for a specific vision task are manually extracted by domain experts, deep learning methods are capable of extracting the feature representations automatically from the raw images or data. Deep learning methods extract high-level

information required for recognition tasks. Deep learning architecture consists of deep neural networks (DNN), which are organised into multi-layered hierarchical trainable layers. As the raw data, given as input to the DNN, flow through each intermediate layer, multiple feature representations are obtained. Each layer learns specific features, with the first several layers learning the low-level features and the deeper layers learning the high-level features. In this manner, the raw data are eventually learned by the multiple layers and are represented by a multi-dimensional feature vector. Thus, deep learning is capable of discriminating different patterns and plays a vital role in scene recognition tasks.

A convolutional neural network [14] (CNN) is one of the commonly used architecture models of deep learning. CNN is a multilayered neural network architecture that mimics the human brain. It consists of convolutional layers, a pooling layer, a non-linear rectified linear unit (ReLU) layer, and a fully connected layer. Convolutional layers are responsible for extracting characteristic features, depending on the number of hidden layers. The pooling layer does the downsampling and reduces the size of feature vectors. The ReLU layer replaces all the negative values with 0 and preserves network stability. The final fully connected layer collects all the high-level features and classifies them as labels. Deep learning gained attention in different areas of computer vision tasks with the success of Alexnet [15], trained on the ImageNet [16] dataset. There are many different popular deep CNN models available for scene recognition tasks, including SqueezeNet [17], VGG16 [18], GoogleNet [19], ResNet [20], DenseNet [21], etc. Some of the commonly used indoor and outdoor datasets are NYU [22], Scene [23], Places [24], SUN [25], MIT 67 [26], etc. The availability of powerful processors and large-scale datasets has improved the efficiency of many deep learning models.

A recent development in deep learning, known as transfer learning, has enhanced the use of deep CNN models in many research applications. Deep models can show their performance when trained with large datasets using powerful GPUs (Graphical Processing Units). But, for some applications, the available dataset is limited. In such situations, we can use the concept of transfer learning to train our deep models. Here, the previously trained deep models transfer their learned parameters to the new learning task, even using a small dataset. The weights of these pretrained models already trained with the large dataset automatically update with the new task of a different dataset. Therefore, recent research has used the concept of transfer learning-based deep models for indoor scene recognition tasks.

In recent years, much research has been conducted in the field of indoor scene recognition. Traditional methods of indoor scene recognition include handcrafted features for classification like texture, colour, etc. In [27], focus is given to these features to classify the images of landscapes and cities. The method worked well for outdoor images, but it faced difficulties in recognising indoor scenes. The authors of [26] solved this problem by combining the global and local features using GIST descriptors and a spatial pyramid of visual words on the MIT-67 dataset. In [28], classification was carried out on the LabelMe dataset by using objects as the feature attribute and SVM as the classifier. In [29], the authors used a probabilistic model using objects as feature representations. They used Adaboost classifiers operated on HOG, grey-level features, and Gabor. Recently, deep learning approaches have been used more than traditional methods. In [30], a deep CNN architecture with a linear SVM classifier was employed on the Scene 15 and MIT-67 datasets and attained accuracies of 90% and 68.24%, respectively. The authors of [31] utilised the methodology of mid-level convolutional features and an SVM classifier on the MIT-67, Scene 15, and NYU datasets and attained accuracies of 74.4%, 93.1%, and 81.2%, respectively. The authors in [32] have employed a novel methodology using scale-invariant and spatial layout convolution activations and an SVM classifier, and they obtained an accuracy of 81.2% on the NYU dataset.

The authors of [33] have proposed a model based on the fusion of transcribed speech with visual and text features using videos of indoor scenes. They achieved accuracies of 70% and 74% on InstaIndoor dataset and YouTubeIndoor dataset, respectively. The researchers

in [34] proposed a novel method using GAN and deep CNN for indoor object detection. They employed the Honey Adam African Vultures Optimisation (HAAVO) algorithm to estimate the optimum distance and attained an accuracy of 94% on my nursing home dataset. The authors of [35] proposed a semantic region relationship model using ResNet50 that employs only semantic segmentation results for indoor scene recognition. They evaluated their performance on the MIT-67, Places365-7, Places365-14, and SUN RGBD datasets with 81.64%, 93.143%, 86.714%, and 76.119% accuracy, respectively. In [36], the authors employed the MRNet using Resnet50 for scene recognition by considering triple information, such as local object information, local scene information, and global scene information. They employed class activation mapping (CAM) to obtain salient features and used LSTM. They achieved accuracies of 96.14%, 88.08%, and 73.98% on Scene 15, MIT 67, and SUN 397, respectively.

The concept of local feature matching was employed by the researchers to overcome some of the challenges of indoor and outdoor scene recognition. The authors of [37] developed a deep-transformer-based network to achieve efficient local feature matching. They also employed a slimming transformer and feature transition module, and they achieved better results by generating robust and accurate matches for both indoor datasets (ScanNet) and outdoor datasets (MegaDepth). The authors of [38] used a detector-free transformer-based CNN model that extracts both global and local features simultaneously. They used the overlapping area prediction module to ensure a clean and effective aggregation of information. Inaccurate match labels were eliminated using the match label weight strategy. The authors of [39] have developed a novel semantic segmentation approach using a matric learning perspective that utilises only a few annotated examples, eliminating the requirement of numerous, densely annotated images. Ref. [40] effectively aggregated the local and global features using an attention-based fusion module. They also used a lightweight feature affine module and mapped the local feature to the normal distribution. They achieved superior performance on the SUN RGB-D and ScanNet V2 datasets. Some of the recent works on scene recognition are compared in Table 1.

Table 1. Some recent existing works on scene recognition.

S. No.	Author	Methodology Used and Results	Merits	Demerits
1	Sitaula et al. [41]	Enhanced VHR attention module (EAM) + atrous spatial pyramid pooling+ global average pooling. Accuracy of 95.39% on AID and 93.04% on NwPU dataset.	Rich discriminative salient features are achieved.	Performance could be improved by using various pretrained models to classify new data.
2	Rafique et al. [42]	Segmentation + feature extraction (Using SegNet, VGG, DCT, DWT) + feature selection using genetic algorithm + neuro fuzzy classifier. Accuracy of 96.13% on Cityscapes, 63.1% for SUN RGB D, and 72.8% on NYU datasets.	Multi-object recognition against any varying environment.	Approach performs well for outdoor scenes when compared to indoor scenes.
3	Yee et al. [43]	CNN + spatial pyramid pooling. Accuracies of 71%, 95.6%, and 98.1% on Event-8, Scene-15, and MIT-67 datasets.	Ensemble learning improves the overall performance.	Data augmentation is employed.
4	Du et al. [44]	TrecgNet and feature selection. Accuracy of 71.8% on NYU depth datasets.	Model performs well with aligned colour and depth information.	Only single modality scene recognition is possible.
5	Ahmed et al. [45]	Fuzzy c-mean, mean shift algorithm + logistic regression classifier. Accuracies of 88.75%, 85.75%, and 80.02% on MSRC, COREL 10K, and CVPR 67 datasets.	Classification of a complex scene is possible,	Performance could be increased by using deep CNN models.
6	Shaopeng et al. [46]	Pretrained ResNet CNN with feature-matching algorithm. Accuracies of 96.49% and 81.69% on scene-15 and MIT 67 datasets.	Eliminates the problem of over fitting.	Performance reduces for images having varying illumination, scale, etc.
7	Romero et al. [47]	Dense SIFT + BoW model + spatial pyramid pooling + binary classifier. Accuracy of 92.64% on ImageCLEF 2012 robot vision dataset.	Could classify scenes of different illumination and scaling.	Slow in execution.

Many of the state-of-the-art methods focus on video-based scene recognition because of the availability of spatial-temporal information from different video frames. Therefore, we have concentrated our work on still image scene recognition. Indoor navigation is a challenging task due to the complexity of the location and position of various objects of the same class. Many research studies on indoor scenes were carried out using image annotation and data augmentation to improve the model's performance. Limited research has been carried out using the NYU depth dataset due to its complexity, clutter, and occluded background. As in many research works, we have not used the deep pretrained model as a black box but have developed a hybrid model incorporating the merits of fuzzy stacking-based preprocessing and K-means clustering-based segmentation to improve our model's performance. The selection of relevant features using World Cup optimisation and the use of deep liquid state machines as classifiers have improved the performance of our model. Considering the aforementioned limitations found in the existing work on scene recognition, we have proposed a fine-tuned pretrained deep CNN hybrid model fused with segmentation and attention modules on the NYU depth dataset that can eliminate the various issues previously mentioned in the related works for the task of indoor scene recognition.

3. Proposed Approach

Our proposed framework for classifying different indoor scenes is shown in Figure 1. We have designed our model using the transfer learning approach to efficiently recognise the different classes of indoor scenes with high accuracy. We have carried this out by implementing fine tuning to the pretrained DenseNet [21] model. We have used the NYU depth dataset [22] in our proposed work. These image samples are first preprocessed using fuzzy colour stacking [48] to remove the noise from the images. These filtered images are then fed to the segmentation module. We then created training and testing samples by dividing the segmented samples (1920) into training images (1536) and testing images (384). The pretrained DenseNet [21] model was trained to extract the image features. Superior features are selected using an adaptive World Cup optimisation algorithm. These relevant features are then fed to a deep LSM (liquid state machine) classifier. The performance of the model is evaluated for accuracy, precision, sensitivity, F1-score, etc. In our previous work [2], we evaluated the performance of Alexnet [15], SqueezeNet [17], ResNet [20], and DenseNet [21] pretrained models for the recognition of human actions. The DenseNet pretrained model showed better performance in terms of accuracy, sensitivity, specificity, and F1-score compared to other models. Therefore, we have preferred the DenseNet [21] pretrained model in this work. Although pretrained CNN attains higher performance on recognition of images that are object-centric, it performs less well when applied directly to complex scenes. This happens due to variation in semantic cues in complex scenes, like indoor scenes. In order to address the problem of spatial variation and obtain regions of interest, we have proposed a segmentation approach before extracting the features by the CNN. Concatenating all the semantic cues for classification may lead the deep model to be less dynamic towards the noisy and redundant details in the indoor scenes. Therefore, we have introduced a feature selection mechanism to ignore irrelevant semantic cues and select the most superior features for classification, and, to a great extent, we can overcome the variation in spatial information in indoor scene recognition. These refined superior features are then fed to a deep LSM (liquid state machine) [49] classifier to classify the indoor scenes.

In this section, we have discussed the workflow of our work, including the feature extraction and classification that we have introduced for the detection of indoor scenes.

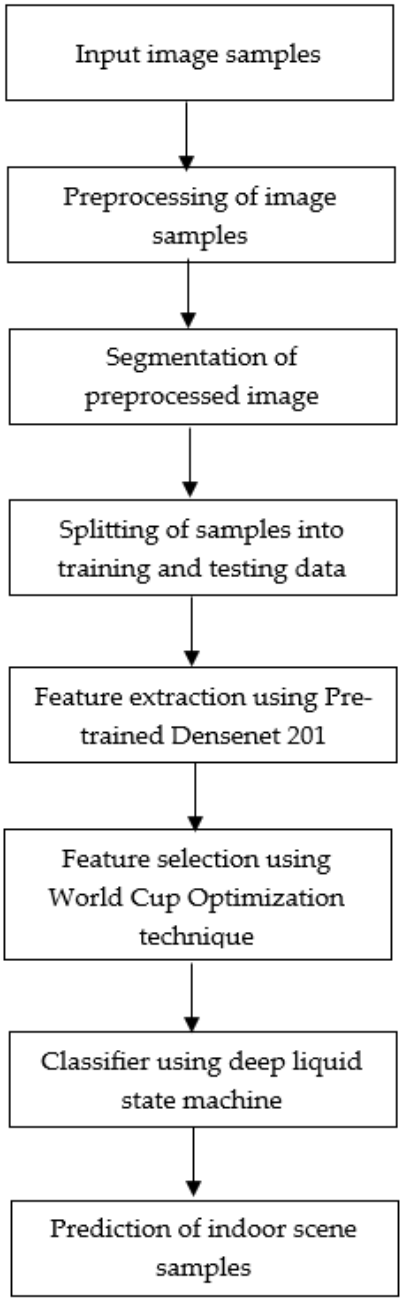

Figure 1. Block diagram of proposed framework.

3.1. Preprocessing: Reconstructing Images Using Fuzzy Colour Technique

The fuzzy technique is accepted because of its high degree of accuracy. The fuzzy colour approach is widely used in image analysis applications. The colour separation is carried out using similarity and difference functions, which are evaluated using membership and non-membership functions. In order to improve the quality of the indoor scenes and eliminate the noisy information, we have preprocessed our dataset using fuzzy colour stacking [48] techniques. They help reduce the noise in the background, thus improving the foreground quality. Here, the input data are separated into a blurred window. The steps involved in the fuzzy technique can be summarised as follows:

- In each window, there is a membership degree associated with each image pixel;
- Based on the distance between the pixel and the window, we calculate the membership degrees;

- We sum up the weights of all the blurred windows, and we create the output image from the average value;
- Two images from a row are combined, and then they are divided into two parts, background and overlay, in order to eliminate the noise from the input image;
- The stacking technique eliminates the noise from the image by considering parameters such as contrast, brightness, opacity, and combining ratio;
- Here, we have stacked our original dataset on the reconstructed dataset using the fuzzy colour technique.

In our work, we have used a contrast value of 1.5, an opacity of 0.6, a brightness value of 80, and a 50% combination ratio. We evaluated our dataset with different values for the above parameters and found that these were more efficient for our dataset. A normalised histogram is evaluated for the input image and the preprocessed image to measure the performance of the preprocessing method.

$$P_r\ (r_k) = n_{k/}n \tag{1}$$

where k = 0 to 255, r_k is the kth intensity level, n_k is the count of pixels with intensity n_k, and n is the total number of pixels in the image.

3.2. Segmentation—Colour Based Approach Using K-Means Clustering

In order to address the problem of spatial variation in indoor scenes, we have proposed a segmentation approach before the CNN extracts the features. Here we aim at segmenting the colour image [50] in the RGB colour model using K-means clustering. The steps involved are described as follows:

- Read the input image;
- Extract the red, green, and blue feature vectors;
- The image space is divided into four group centroids (k = 4);
- Classify the colour pixels using K-means clustering;
- Using the index from K-means clustering, every pixel in the image is labelled;
- Separation of objects in the image by using pixel labels;
- Separate the image by segmenting the cluster centroid.

Thus, all the pixels in each class are identified by the K-means clustering method, and different colours are assigned to each class. The accuracy of the segmentation method is quantitatively evaluated by using the most common matrices, the dice co-efficient (F1-score), and the Jaccard co-efficient (intersection of unity). Both of these matrices are widely used similarity matrices, and they deal with class imbalance.

$$\text{Dice} = \frac{2 * TP}{2 * TP + FP + FN} \tag{2}$$

$$\text{Jaccard} = \frac{TP}{TP + FP + FN} \tag{3}$$

TP—True positive, *FP*—False Positive, *FN*—False Negative.

3.3. Feature Extraction Using Pretrained DenseNet Model

The deep pretrained model, DenseNet (a dense convolutional network) [21], is used in our proposed work to extract the features from the segmented image. Compared to all other CNN models, DenseNet [21] solves the vanishing gradient problem and improves the model's performance. This is achieved by the special architecture of the DenseNet, where each preceding layer concatenates its feature vectors to the future layers, eliminating the effect of reduction in the gradient value propagating through the entire path between the input layer and output layer. DenseNet [21] consists of a special block in its architecture known as a dense block. Each block consists of convolutional layers of sizes 1×1 and

3 × 3. Each dense block is followed by a separate block, known as a transition block, that helps to down sample the feature from the preceding layer.

The architecture of the pretrained DenseNet201 used in our work is shown in Figure 2. In our proposed work, we have removed the fully connected classifier layer from DensNet201, and the features are extracted from the average pool layer. Thus, the architecture of DenseNet201 used in our proposed work consists of a convolutional layer, a maximum pooling layer, a few dense block layers, transition layers, and an average pooling layer. The features are extracted from the average pooling layer of the pretrained DenseNet201. These layers learn the different features from our dataset with better performance, avoiding the vanishing gradient problem.

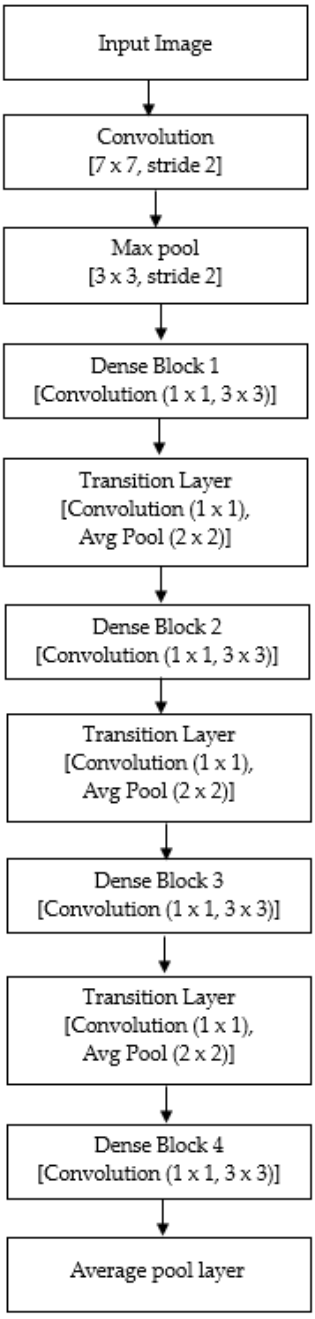

Figure 2. Architecture of DenseNet201.

3.4. World Cup Optimization-Based Feature Selection

One of the important features of the human visual system is its attention capability. Rather than absorbing all the information regarding a context, our attention mechanism selectively chooses the most noticeable features and discards irrelevant information for certain applications. These days, more research has incorporated the attention mechanism in many computer vision tasks to obtain better utilization of the semantic information. In our proposed work, we have embedded the World Cup Optimisation [51] (WCO) method for performing feature selection. Pretrained CNNs extract high-dimensional features from the different images. Considering all this information will increase computational time and consumption. Also, this may reduce the robust nature of the model in noisy and cluttered environments. Thus, discarding irrelevant features will increase the accuracy of the model.

Here, we have the extracted features and the labels, which are fed as input for the feature selection. The weight parameter for each feature attribute is calculated by a fitness function. By using this fitness function, fitness calculations are carried out and the position is updated. The objective of the algorithm is to have a high fitness value. The highest-valued fitness features are selected, and the others are discarded. Finally, selected attribute indices or ranks form the output of the algorithm. These fitness functions and indices, or ranks, are calculated based on the mean and standard deviation. Where 'n' indicates the total number of features of 'X', 'β' lying in the range [0, 1] is the increase or decrease coefficient of 'σ'.

$$\text{Mean, } \mu = \frac{1}{n}\sum_{i=1}^{n} Xi \tag{4}$$

$$\text{Standard Deviation, } \sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} (Xi - \mu)^2} \tag{5}$$

$$\text{Rank} = \frac{(\beta * \sigma + \mu)}{2} \tag{6}$$

3.5. Deep Liquid State Machine Classifier

The superior features of the attention module are given to the classifier module. This is achieved by using a deep LSM (liquid state machine) [49] with an attention module. The deep LSM classifier model, with the architecture shown in Figure 3, is used in our work. It consists of a recurrent and spiking neural network along with multiple read-out neurons. Liquid state machine is a type of deep learning approach that consists of two special layers known as the hidden layer and the winner take all (WTA) layer. LSM consists of three special components: an input layer, a reservoir or liquid layer, and a memoryless readout circuit. The reservoir, or liquid layer, is considered the generic preprocessor that consists of numerous LIF (leaky integrate and fire) neurons. Readout neurons are also known as task processors, and they produce the final output from the LSM.

The probability of synaptic connections between neurons is related to the Euclidean distance between the neurons. The synaptic connection probability P (p, q), from neuron p to q, depends on the Euclidian distance, $D\ (p, q)$, between them:

$$\text{P (p, q)} = C * e^{(-D(p,q)/\tau)^2} \tag{7}$$

where the parameters C and τ regulate the synaptic functions. The scalar parameter C sets the upper limit of the probability, and the parameter τ controls the Euclidean distance between neurons.

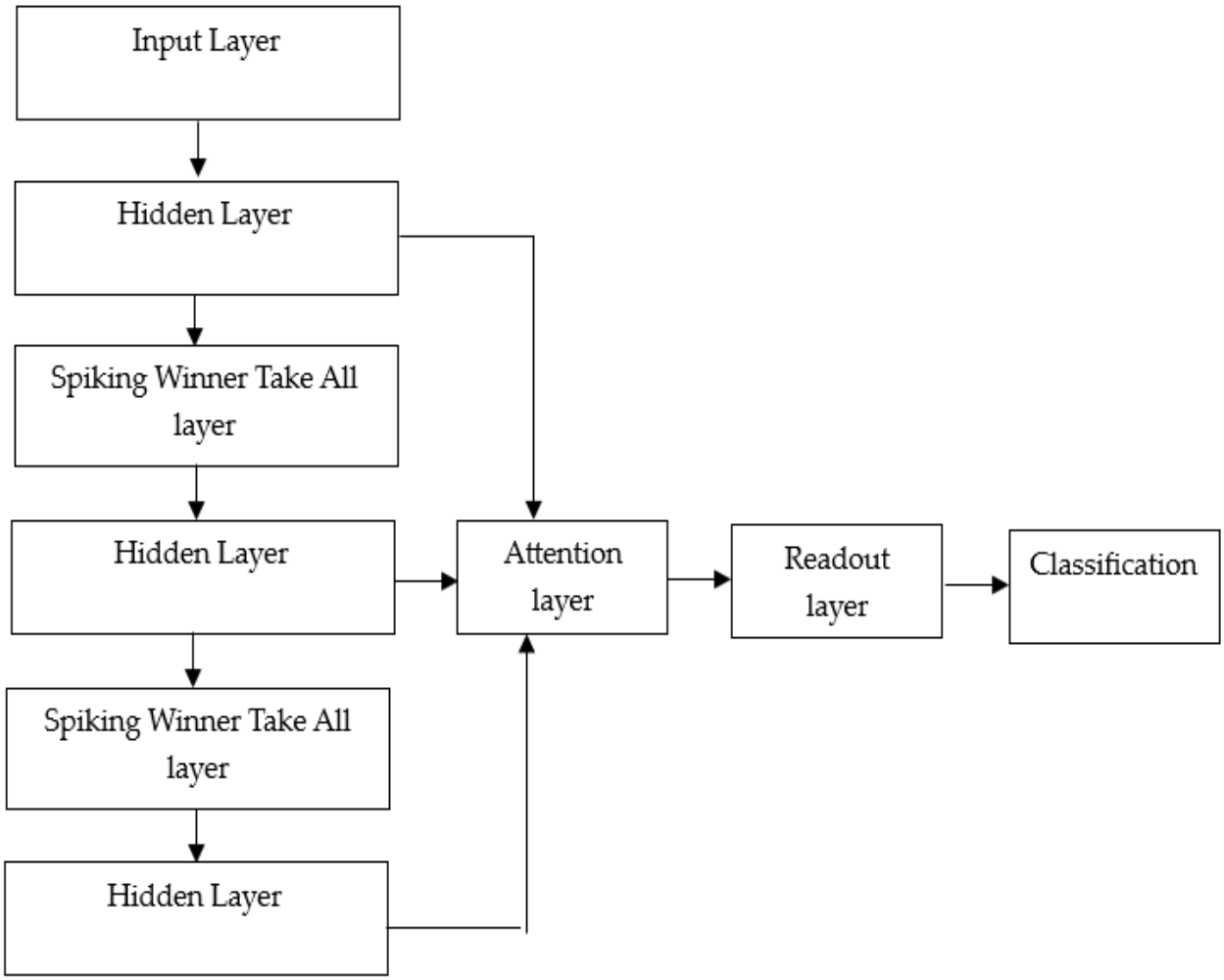

Figure 3. Deep LSM architecture.

4. Dataset

Our proposed work is carried out on the NYU depth dataset [22], https://cs.nyu.edu/~silberman/datasets/nyu_depth_v2.html (accessed on 10 August 2022). It consists of 26 different indoor scene types, consisting of 464 different indoor scenes with 1000+ classes. For our work, we have downloaded 1920 indoor scenes with a total size of 79.5 MB. We have sorted these images, which have varying spatial layouts, orientations, etc., and arranged them into 20 different classes. In our work, we have focused on recognising some specific indoor scene images, such as basement, bathroom, bedroom, bookstore, café, classroom, computer lab, conference room, dinette, dining room, exercise room, furniture store, home office, home storage, kitchen, living room, playroom, reception room, student lounge, and study room. We have collected and sorted the different images in each class based on the diverse nature they exhibit, and the count of images in each class varies accordingly. This could be summarised as having a count of '75', '63', '56', '125', '54', '29', '32', '68', '52', '122', '74', '155', '33', '46', '101', '380', '94', '95', '153', and '110' images in each above-mentioned class, respectively. We have selected these classes with the belief that these indoor scene classes are vital for visually impaired and blind people to navigate safely in their indoor environment. Sample indoor scenes, one from each class, are shown in Figure 4.

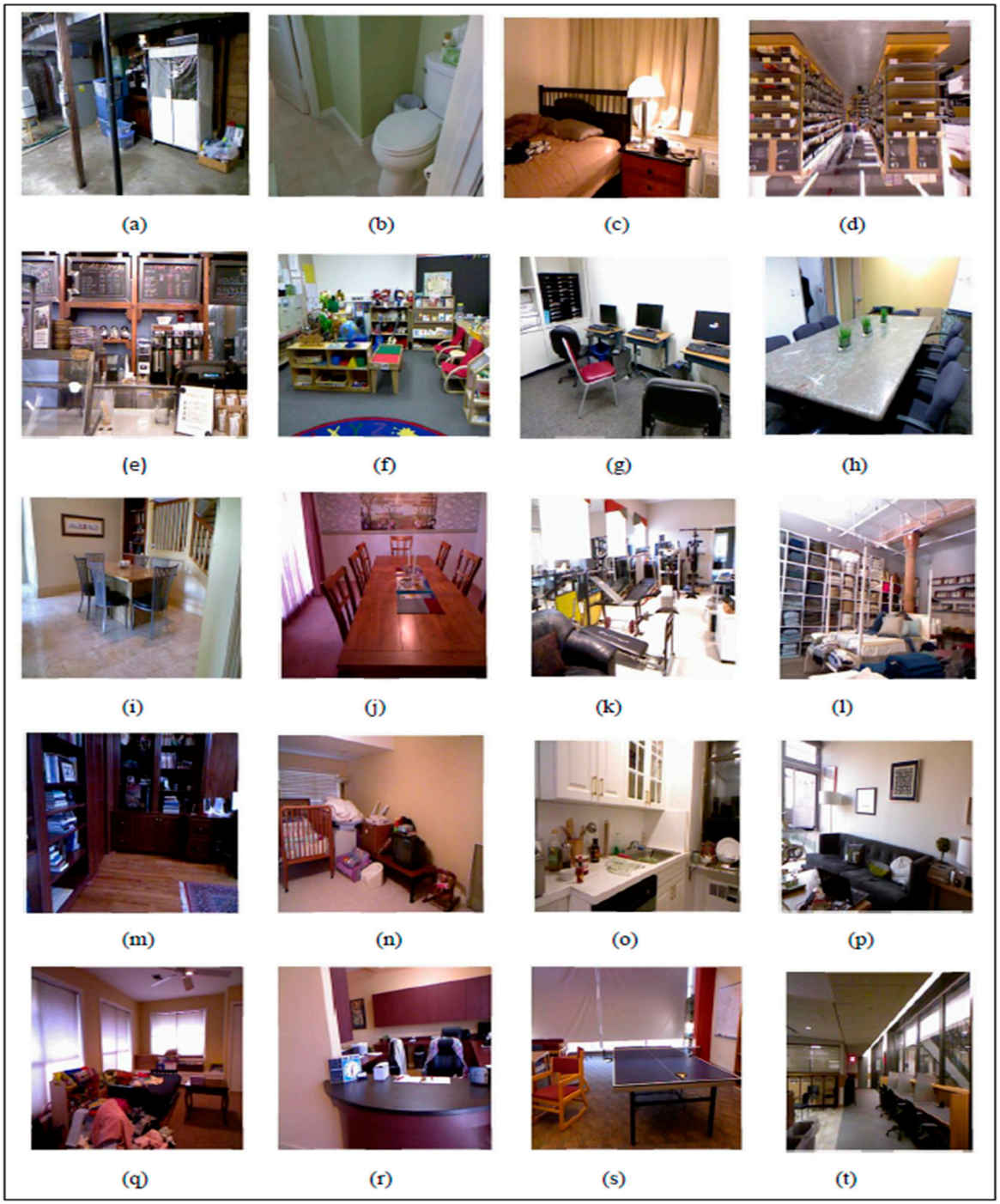

Figure 4. Sample scenes from each class of (**a**) basement, (**b**) bathroom, (**c**) bedroom, (**d**) bookstore, (**e**) cafe, (**f**) classroom, (**g**) computer lab, (**h**) conference room, (**i**) dinette, (**j**) dining room, (**k**) exercise room, (**l**) furniture store, (**m**) home office, (**n**) home storage, (**o**) kitchen, (**p**) living room, (**q**) playroom, (**r**) reception room, (**s**) student lounge, and (**t**) study room.

5. Experimental Results and Discussions

We have conducted our overall experimentation by using the software MATLAB R2021a, a powerful image processing tool. We have used an Intel Core i7 processor with an NVIDIA RTX GPU, a speed of 2.30 GHz, and 16 GB of system RAM.

5.1. Results Analysis

We have evaluated the performance of our proposed model with the NYU depth dataset [22], which consists of numerous categories of different indoor scenes that include scenes of bedrooms, dining rooms, classrooms, conference rooms, etc. These images were preprocessed using fuzzy colour stacking to remove the background noise. These preprocessed images are then segmented using colour-based segmentation techniques and K-means clustering to solve the issues of spatial variance and layout. These segmented images are then fed to a pretrained DenseNet for feature extraction. These extracted features are then given to an attention module utilising World Cup optimisation for feature selection. These selected superior features are classified using a deep LSM (liquid state machine) classifier [49]. Figure 5 shows the input scene of a bedroom image and the corresponding preprocessed image. The foreground picture quality is improved by reducing the background noise using the fuzzy colour stacking technique [48]. Figure 6 shows the normalised histogram plot for the sample input image and its corresponding stacked image.

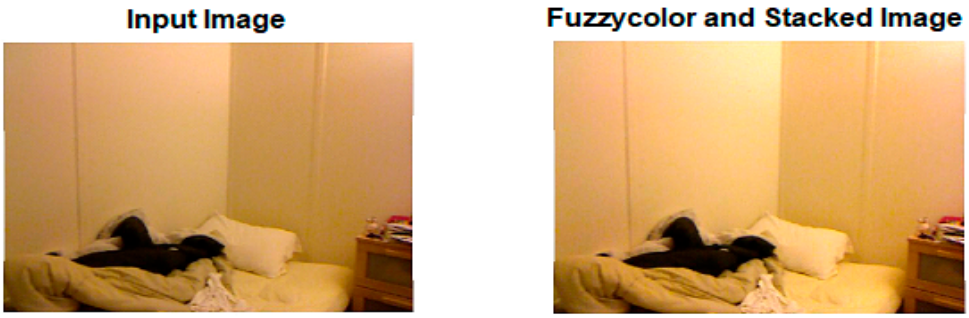

Figure 5. Sample input image of a bedroom and its corresponding preprocessed image.

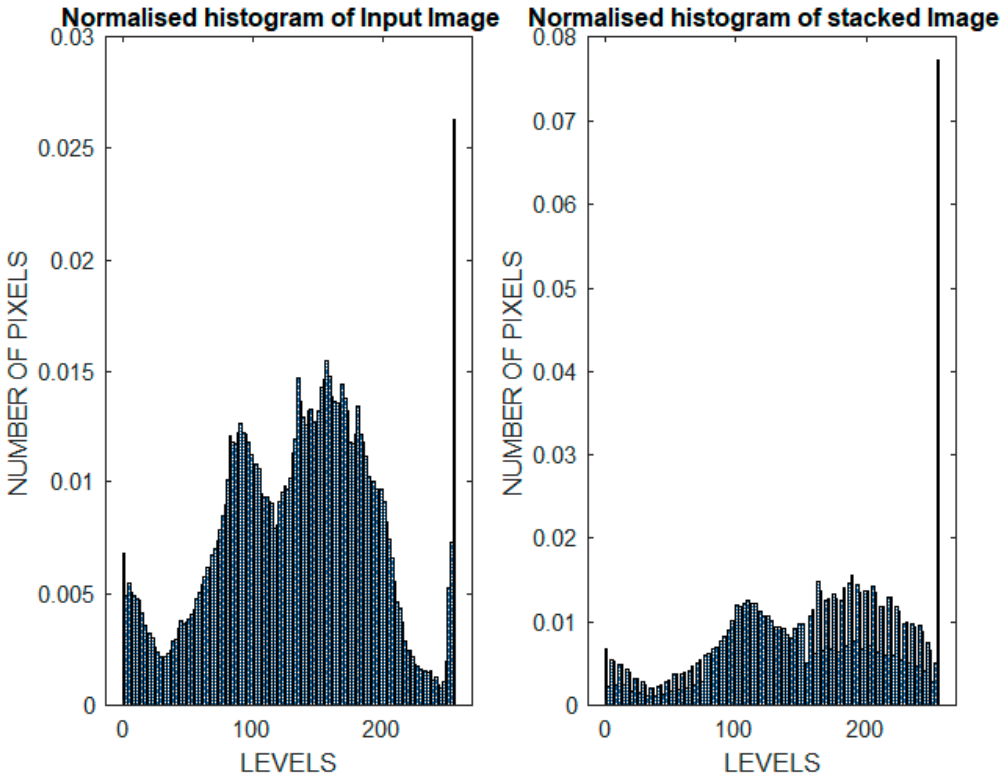

Figure 6. Normalized histogram plot for sample input image of a bedroom and its corresponding preprocessed image.

Figure 7 shows the segmented results of some sample images, preserving the semantic cues and spatial information of our sample indoor scene.

Figure 7. Sample images of (**a**) bedroom, (**b**) bathroom, (**c**) dining room, and (**d**) kitchen with input image (**left**) and segmented image (**right**).

We have also evaluated the accuracy of our segmentation module by performing a pixel-wise comparison between our segmented image and its corresponding ground truth. We have estimated the similarity measure by calculating the dice coefficient (F1 score) and Jaccard's index (intersection over union). Figure 8 shows the plot of the dice coefficient, and Figure 9 shows the Jaccard's coefficient. Both plots show a maximum co-efficient value of 1, which indicates good accuracy for our segmentation module.

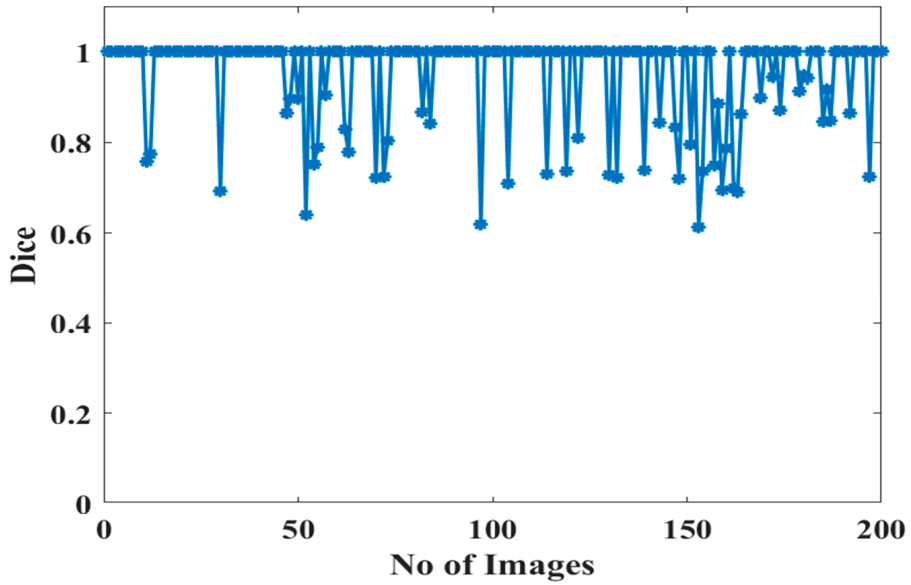

Figure 8. Plot of Dice coefficient of the segmented images and their ground truth.

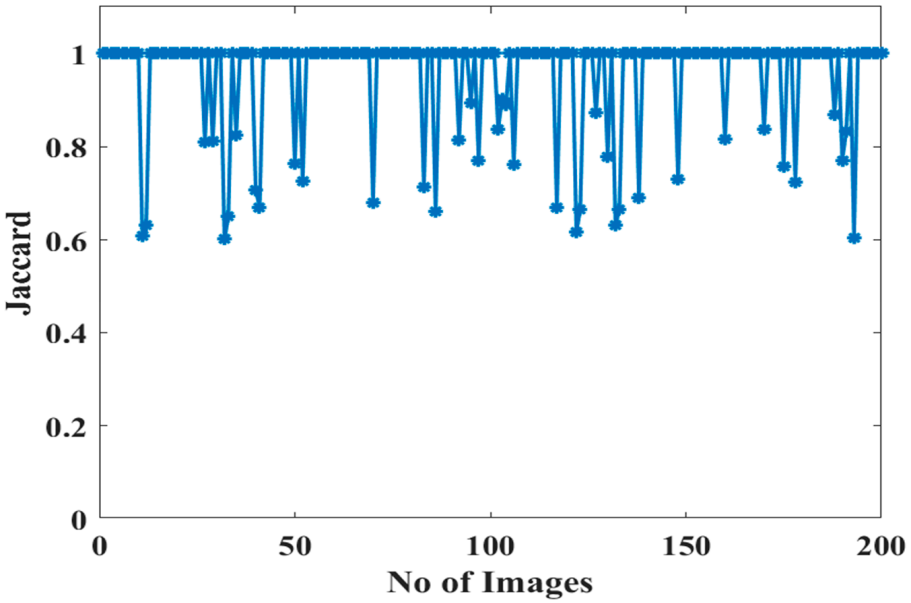

Figure 9. Plot of Jaccard's index of the segmented images and their ground truth.

The dataset used in the proposed model consists of a total of 1920 images that belong to 20 different indoor classes. We have divided 80% of the dataset into training sets and 20% into testing phases. Therefore, we have used 1536 images for the training phase and 384 images for the testing phase. Out of the 1536 images in the training set, we used 20% for validation, and the remaining images were used for training. We have trained our model using the Adam optimizer with the gradient threshold set to '1'. The model was trained with a learning rate of '0.01' that drops by a factor of '0.2'. We have trained and evaluated the performance of our model using a mini batch size of '128' for 500 iterations. Figure 10 shows the plots of accuracy and loss for the training and validation processes of our proposed model.

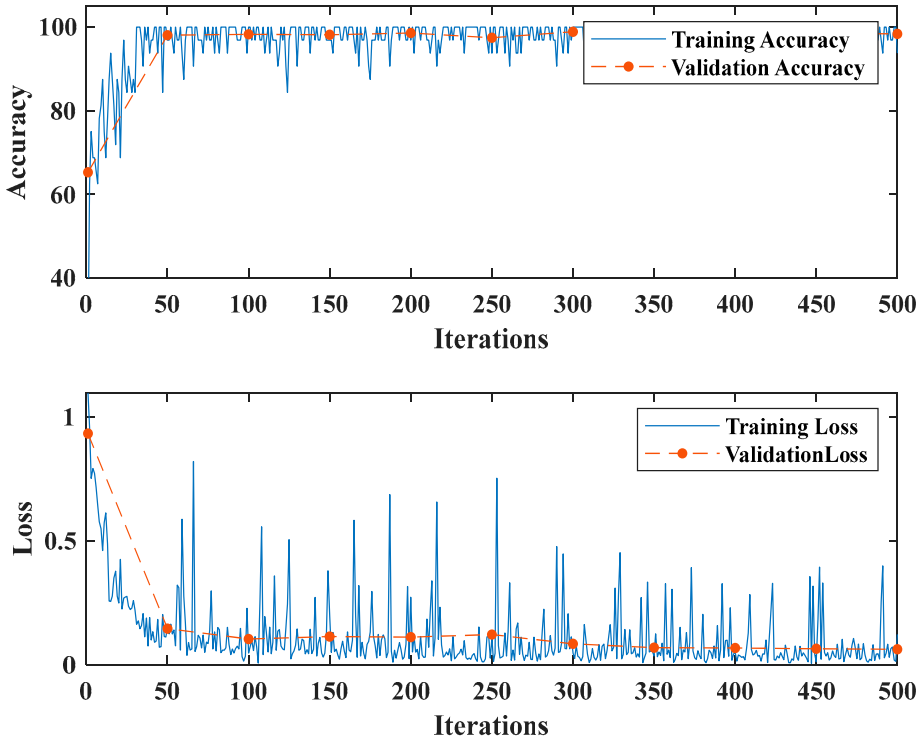

Figure 10. Accuracy and loss plots of training and validation.

5.2. Classification Report–Performance Evaluation

The performance evaluation of our proposed indoor scene recognition model is represented by using matrices such as accuracy, sensitivity, specificity, precision, and F1-score, as shown in Table 2, and the ROC plot for our model is shown in Figure 11.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{8}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{9}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{10}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{11}$$

$$\text{F1} - \text{score} = \frac{2 * recall * precision}{recall + precision} \tag{12}$$

TP—True positive, *TN*—True Negative, *FP*—False Positive, *FN*—False Negative

Table 2. Performance matrices of our proposed model for indoor scene classification.

Proposed Model	Specificity	Sensitivity	Precision	F1-Score
DenseNet201 (Feature extraction) + LSM classifier	0.96	1	0.94	0.95
Accuracy		0.96%		

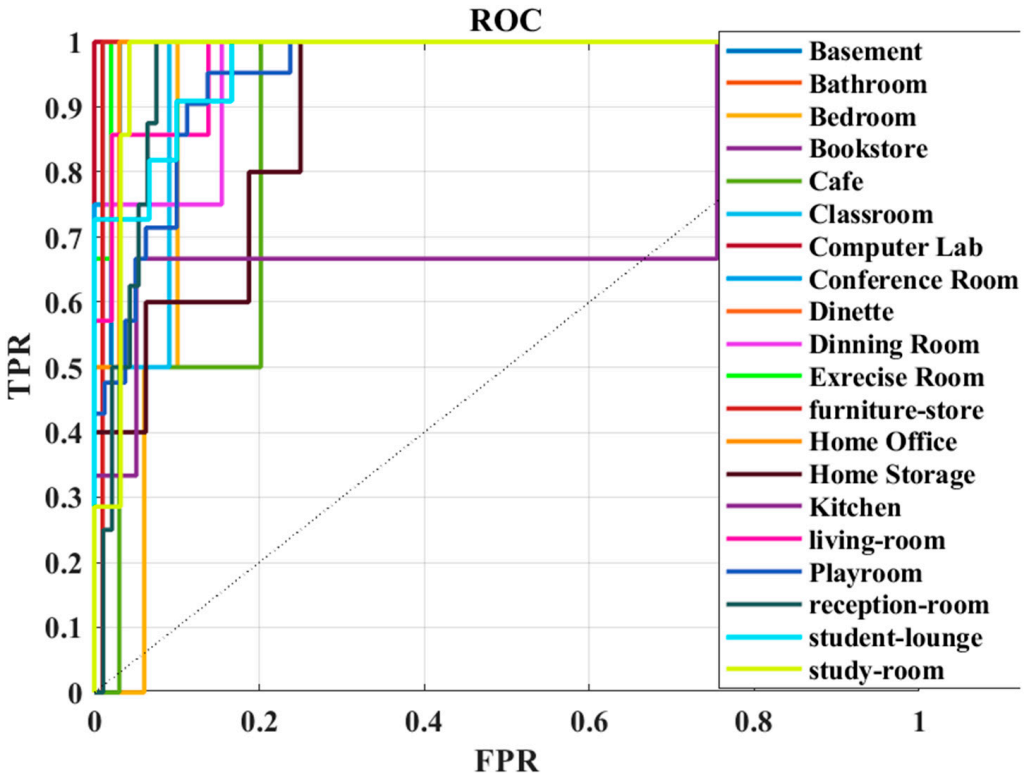

Figure 11. ROC curve.

5.3. Discussions

As a part of indoor scene recognition, we have proposed a hybrid model that combines the features of deep learning and fuzzy logic, segmentation, and feature selection to predict and recognise indoor scenes in the NYU depth dataset with high accuracy. Here, we have experimented on 20 different classes of the NYU depth dataset that are considered vital in assisting visually handicapped and elderly people in their indoor navigation. The image dataset is preprocessed using fuzzy colour stacking to remove the noise from the background and enhance the foreground details. These preprocessed images are then segmented to preserve the semantic cues and localization. Then, we used a pretrained DenseNet201 deep model for feature extraction. The extracted features were filtered, and only the predominant features were selected using the World Cup optimisation algorithm. The removal of irrelevant features improves the robust nature of the model. A deep liquid state machine trained on these selected features predicts our indoor scene classes with good accuracy. Our model showcases a specificity of 96%, a sensitivity of 100%, a precision of 94%, and an F1-score of 95%. By evaluating the performance matrices, we can conclude that, for the given indoor dataset, our DenseNet-LSM model obtained an accuracy of 96%.

6. Comparison of Our Proposed Work to Existing Indoor Scene Recognition Research

Many recent research studies have employed different advanced deep learning and machine learning models to perform scene recognition. We have compared our result with some works related to indoor scene recognition. In this section, we have compared the research works on the NYU depth dataset. Although our comparison findings may not be truly equitable, since many research works use different datasets, they offer an insight into the various classification methods and their outcomes. Table 3 shows the comparison of the accuracy of our proposed model with other existing research studies for indoor scene recognition.

Table 3. Accuracy comparison of existing works with our proposed model on indoor scene recognition.

S. No.	Author	Methodology Used	Accuracy
1	Proposed Model	Segmentation + DenseNet201 + World Cup Optimization + LSM Classifier	96%
2	Pereira et al. [52]	Semantic Segmentation + VGG16, ResNet18-50-101, DenseNet and MobileNetV2+ Feature fusion	75.8%
3	Heikel et al. [53]	YOLO + TF + IDF	83.63%
4	Mosella et al. [54]	2D–3D geometric feature fusion + Graph convolutional neural network	75%
5	Afif et al. [55]	EfficientNet CNN model + Scaling	95.6%
6	Li et al. [56]	MAPNet + Attentive pooling	67.7%
7	Guo et al. [57]	GoogleNet + Inception V3 + Feature fusion	96%
8	Tang et al. [58]	GoogleNet + Multi feature fusion	92.92%

7. Conclusions and Future Scope

In this paper, we have proposed a segmentation-based attention model for indoor scene recognition that can assist visually impaired and elderly people with indoor navigation. Here, we have used the transfer learning concept of deep learning to develop our framework. We have implemented our work using a deep-pretrained DenseNet201 CNN architecture and a deep LSM (liquid state machine) model. We have evaluated our model on the NYU depth dataset of 20 classes of indoor scenes. We have tried to improve the robustness and performance of our model in varying indoor scenes by combining the advantages of a few techniques like fuzzy colour stacking, segmentation, and World Cup optimization. Preprocessing by the fuzzy colour stacking technique has helped improve the foreground quality of the image dataset. By adding the segmentation module, our model was able to handle the spatial details and semantic cues, providing a better region of interest in the indoor scenes. The pretrained DenseNet201 extracted the features from these segmented images. These features were filtered, and the most predominant features were selected using the World Cup optimisation algorithm. We have used a deep LSM (liquid state machine) model as our classifier, which efficiently classified the 20 classes of our indoor dataset. Our proposed model could achieve an accuracy of 96% on the NYU dataset. Thus, we could improve the robustness of our proposed model to recognise the different indoor scenes. In our future work, we will emphasise working with different datasets, including outdoor scenes, and combining various modalities in order to expand the relevance and applicability of our proposed work.

Author Contributions: Conceptualization, R.S.; Methodology, I.C.; Validation, D.J.H.; Formal analysis, J.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Herranz, L.; Jiang, S.; Li, X. Scene recognition with CNNs: Objects, scales and dataset bias. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 571–579.
2. Surendran, R.; Anitha, J.; Hemanth, J.D. Recognition of human action for scene understanding using world cup optimization and transfer learning approach. *PeerJ Comput. Sci.* **2023**, *9*, e1396. [CrossRef] [PubMed]
3. Hernandez, A.C.; Gomez, C.; Barber, R.; Mozos, O.M. Exploiting the confusions of semantic places to improve service robotic tasks in indoor environments. *Robot. Auton. Syst.* **2023**, *159*, 104290. [CrossRef]
4. Guo, J.; Nie, X.; Ma, Y.; Shaheed, K.; Ullah, I.; Yin, Y. Attention based consistent semantic learning for micro-video scene recognition. *Inf. Sci.* **2021**, *543*, 504–516. [CrossRef]
5. Bosch, A.; Muñoz, X.; Martí, R. Which is the best way to organize/classify images by content. *Image Vis. Comput.* **2007**, *25*, 778–791. [CrossRef]
6. Brown, M.; Susstrun, S.K. Multi-spectral SIFT for scene category recognition. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 20–25 June 2011; pp. 177–184.
7. Bay, H.; Ess, A.; Tuytelaars, T.; van Gool, L. Speeded-up robust features (SURF). *Comput. Vis. Image Underst.* **2008**, *110*, 346–359. [CrossRef]
8. Oliva, A.; Torralba, A. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vis.* **2001**, *42*, 145–175. [CrossRef]
9. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–26 June 2005; Volume 1, pp. 886–893.
10. Yang, J.; Jiang, Y.G.; Hauptmann, A.; Ngo, C.W. Evaluating bag-of-visual-words representations in scene classification. In Proceedings of the International Workshop on Multimedia Information Retrieval, Bavaria, Germany, 24–29 September 2007; IEEE: Augsburg, Germany, 2007; pp. 197–206.
11. Sánchez, J.; Perronnin, F.; Mensink, T.; Verbeek, J. Image classification with the fisher vector: Theory and practice. *Int. J. Comput. Vis.* **2013**, *105*, 222–245. [CrossRef]
12. Jégou, H.; Douze, M.; Schmid, C.; Pérez, P. Aggregating local descriptors into a compact image representation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 3304–3311.
13. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 7553. [CrossRef]
14. Lecun, Y.; Boser, B.; Denker, J.S.; Henderson, D.; Howard, R.E.; Hubbard, W.; Jackel, L.D. Backpropagation applied to handwritten zip code recognition. *Neural Comput.* **1989**, *1*, 541–551. [CrossRef]
15. Krizhevsky, I.; Sutskever, G.E. Hinton. Imagenet classification with deep convolutional neural networks. *Proc. Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105.
16. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
17. Iandola, F.N.; Han, S.; Moskewicz, M.W.; Ashraf, K.; Dally, W.J.; Keutzer, K. 2016 SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size. *arXiv* **2016**, arXiv:1602.07360.
18. Simonyan, K.; Zisserman, A. 2015 Very deep convolutional networks for large-scale image recognition, ICLR. *arXiv* **2014**, arXiv:1409.1556.
19. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
20. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
21. Huang, G.; Liu, Z.; Der Maaten, L.V.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269.
22. Silberman, N.; Hoiem, D.; Kohli, P.; Fergus, R. Indoor segmentation and support inference from RGB-D images. In Proceedings of the 12th European Conference on Computer Vision (ECCV), Florence, Italy, 7–13 October 2012; pp. 746–760.
23. Lazebnik, S.; Schmid, C.; Ponce, J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), New York, NY, USA, 17–22 June 2006; Volume 2, pp. 2169–2178.
24. Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; Torralba, A. Places: A 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 1452–1464. [CrossRef]
25. Xiao, J.; Hays, J.; Ehinger, K.A.; Oliva, A.; Torralba, A. Sun database: Large-scale scene recognition from abbey to zoo. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA, 13–18 June 2010; pp. 3485–3492.
26. Quattoni, A.; Torralba, A. Recognizing indoor scenes. In Proceedings of the IEEE Conference on Computer and Pattern Recognition (CVPR), Miami, FL, USA, 20–25 June 2009; pp. 413–420.
27. Vailaya, A.; Figueiredo, M.A.T.; Jain, A.K.; Zhang, H.-J. Image classification for content-based indexing. *IEEE Trans. Image Process.* **2001**, *10*, 117–130. [CrossRef] [PubMed]

28. Li, L.J.; Su, H.; Lim, Y.; Fei-Fei, L. Objects as attributes for scene classification. In Proceedings of the European Conference on Computer Vision, Heraklion, Crete, 5–11 September 2010; Springer: Berlin/Heidelberg, Germany; pp. 57–69.
29. Espinace, P.; Kollar, T.; Soto, A.; Roy, N. Indoor scene recognition through object detection. In Proceedings of the IEEE International Conference on Robotics and Automation, Anchorage, Alaska, 3–8 May 2010.
30. Zhou, B.; Lapedriza, A.; Xiao, J.; Torralba, A.; Oliva, A. Learning deep features for scene recognition using places database. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 487–495.
31. Khan, S.H.; Hayat, M.; Bennamoun, M.; Togneri, R.; Sohel, F.A. A discriminative representation of convolutional features for indoor scene recognition. *IEEE Trans. Image Process.* **2016**, *25*, 3372–3383. [CrossRef] [PubMed]
32. Hayat, M.; Khan, S.H.; Bennamoun, M.; An, S. A spatial layout and scale invariant feature representation for indoor scene classification. *IEEE Trans. Image Process.* **2016**, *25*, 4829–4841. [CrossRef]
33. Glavan, A.; Talavera, E. InstaIndoor and multi-modal deep learning for indoor scene recognition. *Neural Comput. Appl.* **2022**, *34*, 6861–6877. [CrossRef]
34. Nagarajan, A.; Gopinath, M.P. Hybrid Optimization-Enabled Deep Learning for Indoor Object Detection and Distance Estimation to Assist Visually Impaired Persons. *Adv. Eng. Softw.* **2023**, *176*, 103362. [CrossRef]
35. Song, C.; Ma, X. SRRM: Semantic Region Relation Model for Indoor Scene Recognition. *arXiv* **2023**, arXiv:2305.08540.
36. Lin, C.; Lee, F.; Xie, L.; Cai, J.; Chen, H.; Liu, L.; Chen, Q. Scene recognition using multiple representation network. *Appl. Soft Comput.* **2022**, *118*, 108530. [CrossRef]
37. Xie, T.; Dai, K.; Wang, K.; Li, R.; Zhao, L. Deepmatcher: A deep transformer-based network for robust and accurate local feature matching. *arXiv* **2023**, arXiv:2301.02993. [CrossRef]
38. Dai, K.; Xie, T.; Wang, K.; Jiang, Z.; Li, R.; Zhao, L. OAMatcher: An Overlapping Areas-based Network for Accurate Local Feature Matching. *arXiv* **2023**, arXiv:2302.05846.
39. Wang, K.; Liew, J.H.; Zou, Y.; Zhou, D.; Feng, J. Panet: Few-shot image semantic segmentation with prototype alignment. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9197–9206.
40. Xie, T.; Wang, L.; Li, R.; Zhang, X.; Zhang, H.; Yang, L.; Liu, H.; Li, J. FARP-Net: Local-Global Feature Aggregation and Relation-Aware Proposals for 3D Object Detection. *IEEE Trans. Multimed.* **2023**, 1–15. [CrossRef]
41. Sitaula, C.; KC, S.; Aryal, J. Enhanced Multi-Level Features for Very High-Resolution Remote Sensing Scene Classification. *arXiv* **2023**, arXiv:2305.00679.
42. Rafique, A.A.; Ghadi, Y.Y.; Alsuhibany, S.A.; Chelloug, S.A.; Jalal, A.; Park, J. CNN Based Multi-Object Segmentation and Feature Fusion for Scene Recognition. In Proceedings of the Conference on Membrane Computing, Chandler, AZ, USA, 27–29 April 2022.
43. Yee, P.S.; Lim, K.M.; Lee, C.P. DeepScene: Scene classification via convolutional neural network with spatial pyramid pooling. *Expert Syst. Appl.* **2022**, *193*, 116382. [CrossRef]
44. Du, D.; Wang, L.; Li, Z.; Wu, G. Cross-modal pyramid translation for RGB-D scene recognition. *Int. J. Comput. Vis.* **2021**, *129*, 2309–2327. [CrossRef]
45. Ahmed, A.; Jalal, A.; Kim, K. A Novel Statistical Method for Scene Classification Based on Multi-Object Categorization and Logistic Regression. *Sensors* **2020**, *20*, 3871. [CrossRef] [PubMed]
46. Liu, S.; Tian, G. An Indoor Scene Classification Method for Service Robot Based on CNN Feature. *J. Robot.* **2019**, *2019*, 8591035. [CrossRef]
47. Romero-González, C.; Martínez-Gómez, J.; García-Varea, I.; Rodriguez-Ruiz, L. On robot indoor scene classification based on descriptor quality and efficiency. *Expert Syst. Appl.* **2017**, *79*, 181–193. [CrossRef]
48. Toğaçar, M.; Ergen, B.; Cömert, Z. COVID-19 detection using deep learning models to exploit Social Mimic Optimization and structured chest X-ray images using fuzzy color and stacking approaches. *Comput. Biol. Med.* **2020**, *121*, 103805. [CrossRef]
49. Oladipupo, G.G. Research on the Concept of Liquid State Machine. *arXiv* **2019**, arXiv:1910.03354.
50. Chitade, A.Z.; Katiyar, S.K. Colour based image segmentation using k-means clustering. *Int. J. Eng. Sci. Technol.* **2010**, *2*, 5319–5325.
51. Razmjooy, N.; Khalilpour, M.; Ramezani, M. A New Meta-Heuristic Optimization Algorithm Inspired by FIFA World Cup Competitions: Theory and Its Application in PID Designing for AVR System. *J. Control. Autom. Electr. Syst.* **2016**, *27*, 419–440. [CrossRef]
52. Pereira, R.; Barros, T.; Garrote, L.; Lopes, A.; Nunes, U.J. A Deep Learning-based Global and Segmentation-based Semantic Feature Fusion Approach for Indoor Scene Classification. *arXiv* **2023**, arXiv:2302.06432.
53. Heikel, E.; Espinosa-Leal, L. Indoor Scene Recognition via Object Detection and TF-IDF. *J. Imaging* **2022**, *8*, 209. [CrossRef] [PubMed]
54. Mosella-Montoro, A.; Ruiz-Hidalgo, J. 2d–3d geometric fusion network using multi-neighbourhood graph convolution for rgb-d indoor scene classification. *Inf. Fusion* **2021**, *76*, 46–54. [CrossRef]
55. Afif, M.; Ayachi, R.; Said, Y.; Atri, M. Deep learning-based application for indoor scene recognition. *Neural Process. Lett.* **2020**, *51*, 2827–2837. [CrossRef]
56. Li, Y.; Zhang, Z.; Cheng, Y.; Wang, L.; Tan, T. MAPNet: Multi-modal attentive pooling network for RGB-D indoor scene classification. *Pattern Recognit.* **2019**, *90*, 436–449. [CrossRef]

57. Guo, W.; Wu, R.; Chen, Y.; Zhu, X. Deep learning scene recognition method based on localization enhancement. *Sensors* **2018**, *18*, 3376. [CrossRef]
58. Tang, P.; Wang, H.; Kwong, S. G-MS2F: GoogLeNet based multi-stage feature fusion of deep CNN for scene recognition. *Neurocomputing* **2017**, *225*, 188–197. [CrossRef]