

Confidence-driven feature selection and ensemble classification using ResNet, AlexNet, and DenseNet for scene recognition

Sumathi Krishnamurthy¹, Shetty Pramod Kumar¹, Halasur Rudrappa Mahadevaswamy²,
Ujwala Bagepalli Satishbabu¹

¹Department of Electronics and Communication Engineering, Jawaharlal Nehru New College of Engineering,
Visvesvaraya Technological University, Belagavi, India

²Jagadguru Sri Shivarathreshwara University, Noida, India

Article Info

Article history:

Received Oct 23, 2025

Revised Jul 17, 2026

Accepted Aug 12, 2026

Keywords:

Confidence-driven world cup
optimization

Contrast-limited adaptive
histogram equalization

Deep liquid state machine
classifier

Feature extraction

Feature selection

Retinex

ABSTRACT

Indoor scene recognition is a significant problem because of its structural complexity, illumination variations, and visual similarities between the various scene classes. Recent approaches use convolutional neural networks (CNNs) and transfer learning to extract features for classification. In these methods, all the features were considered for classification, which leads to an increase in computational complexity and training time. This leads to more time consumption for processing. To address these challenges, this study proposed a hybrid indoor scene recognition framework that integrates illumination-aware pre-processing, multi-model deep feature extraction, and confidence-based feature fusion. The pre-processing technique uses contrast-limited adaptive histogram equalization (CLAHE) and Retinex enhancement. Deep features are extracted using Alex Krizhevsky network (AlexNet), densely connected convolutional network (DenseNet), and residual network 101 layers (ResNet101) to capture spatial and semantic information. The features extracted are fed to confidence-driven world cup optimization (CD-WCO) to extract the significant features. The final representation is classified using a deep liquid state machine (DLSM) classifier to enhance class discrimination. Our experiments, conducted on the NYU indoor scene dataset with 20 scene categories and the MIT Indoor-15 class subset results indicate the effectiveness of the proposed framework, achieving an overall classification accuracy of 96.5% and 95.2%, respectively.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Sumathi Krishnamurthy

Department of Electronics and Communication Engineering, Jawaharlal Nehru New College of Engineering

Visvesvaraya Technological University

Belagavi, India

Email: sumathik@jnnce.ac.in

1. INTRODUCTION

Scene recognition is a high-level computer vision task that involves classifying an image or video based on its overall environment and context, such as identifying it as a “beach,” “kitchen,” or “street” [1]. Scene recognition not only concentrates on identifying the class of the image, but also focuses on the semantic meaning associated with the particular scene [2]. Scene identification has a wide range of applications in robotics, autonomous driving, and medical imaging. to achieve the scene classification task, the steps are feature detection, feature description, and classification [3]. Scene recognition using convolutional neural network (CNN) and transfer learning-based techniques has significantly increased the

efficiency of scene classification. The scene can be static or dynamic. Dynamic scenes are referred to as video sequences, and static scenes are still images [4]. Static scene analysis is dependent on single-frame information. Static scene classification is further divided into indoor and outdoor scene classification. Outdoor scenes, such as streets, forests, or beaches—generally contain similar object sets, while indoor scenes—such as classrooms, bedrooms, or living rooms—feature a variety of objects with differing dimensions, spatial arrangements, and levels of occlusion. Indoor scene classification plays a very important role in several computer applications, including robotics navigation, human-computer interaction, and content-aware systems. Unlike outdoor scene images, indoor scenes have high intra-class variability, strong illumination changes, and the presence of visually similar objects across different categories. The above challenges make indoor scene classification a non-trivial task, even with recent advances in deep learning.

Early approaches to scene recognition relied on handcrafted features, such as scale-invariant feature transform (SIFT), histogram of oriented gradients (HOG), and global image descriptor (GIST) descriptors [2]. The above approaches capture local texture and structural information; they are highly sensitive to illumination variations. These handcrafted features require manual design and lack the representational capacity to generalize across complex indoor scene images, resulting in limited classification accuracy. With the evolution of deep learning, scene recognition using CNN and transfer learning-based techniques has significantly increased the efficiency of scene classification. CNNs became the dominant approach for scene classification due to their ability to learn hierarchical feature representations. Architectures such as residual network (ResNet), Alex Krizhevsky network (AlexNet), and densely connected convolutional network (DenseNet) have shown exceptional performance in extracting features from image data, which aids in classifying scene images [5]. However, despite the exceptional performance from these models, challenges remain in selecting diversified and distinguished features from the entire feature set extracted from the deep learning models and combining the diverse features from the multiple models.

However, single-model CNN-based approaches have several limitations. First, they often overfit to dominant visual patterns, leading to poor generalization across various indoor scenes. Second, deeper networks increase computational complexity and training time, making them less suitable for resource-constrained environments. Recent advancements in feature fusion and ensemble learning techniques have addressed these drawbacks of combining features from multiple deep architectures. The multi-model fusion improves classification accuracy because of high representational diversity, employing simple concatenation or fixed-weight averaging. Such static fusion mechanisms fail to adapt to class-specific confidence variations and may introduce redundant or noisy features, ultimately limiting performance gains. To overcome this limitation, in our paper, we have proposed a novel method of a hybrid scene recognition framework combining preprocessing techniques such as Retinex and contrast-limited adaptive histogram equalization (CLAHE) [6] for illumination normalization with confidence-driven world cup optimization (CD-WCO) for diversified feature selection and a deep liquid state machine (DLSM) classifier at the classification stage, with a spiking winner-takes-all (SWTA) [7] layer that enhances non-linear decision boundaries, to classify the scene images, resulting in improved illumination robustness and scene discrimination performance.

2. LITERATURE REVIEW

Over the years, many deep learning approaches have been introduced in scene recognition. architectures such as ResNet, AlexNet, and DenseNet have shown exceptional performance in extracting features from image data, which aids in classifying scene images. However, despite the exceptional performance from these models, challenges remain in selecting diversified and distinguished features from the entire feature set extracted from the deep learning models and combining the diverse features from the multiple models [5]. To address these challenges, in this research work proposed an attention mechanism that acts as a powerful tool in focusing on the most important features representing the informative data of the images, and in parallel, ensemble learning techniques combine the strengths of multiple models, leveraging their diverse feature extraction capabilities to achieve more efficient performance than any single model alone. In this proposed research work introduced a confidence-modulated world cup optimization (CM-WCO) [8] technique for feature selection and an ensemble learning framework that integrates ResNet, AlexNet, and DenseNet for robust scene recognition.

This research work concentrates on contributing to the development of more efficient and accurate scene recognition systems by combining advanced deep learning architectures with attention-guided feature selection and ensemble strategies, providing a way for improved scene recognition. Scene classification is a fundamental task within the broader field of scene understanding, where the objective is to interpret a given scene and assign it to a specific category. The early recognition process was completely dependent on pixel-based representations, where the raw pixels of the scene were used for classification. These methods

often faced limitations related to image resolution and the granularity of pixel information. To overcome this limitation, researchers have shifted to object-based classification, which utilizes object-based information to classify the scene images. But object-based classification does not concentrate on the semantic context of the scene images. This challenge has led researchers to move towards the semantic understanding of scene images. Scene classification involving the semantic information depends on identifying the distinct features of the image. Some methods, such as SIFT [9] and HOG, are based on handcrafted feature techniques that were used to classify images, but as these techniques rely on manually designed filters, they are prone to missing important feature details, limiting their effectiveness. Additional feature extraction techniques have been developed, such as spatial pyramid matching and fisher vector coding. Once these features have been extracted, they have been fed into the machine learning algorithms like k-means clustering, sparse coding, and principal component analysis (PCA) [10] to perform the classification function. The efficiency of these methods is dependent on how the features are fed into the model.

The advent of the deep learning concept of extracting high-level features in the given dataset has paved way new era has been paved. Deep learning models extract the features automatically from the data. CNNs are used to extract the features automatically. Major developments in the domain of deep learning happened after the development of the AlexNet network, which is pre-trained on the ImageNet dataset. Many deep learning models have been developed, which include visual geometry group network (VGGNet) [11], GoogLeNet [12], squeeze network (SqueezeNet), ResNet, and DenseNet. With the advent of large image datasets and powerful processors, many research works are being carried out with these deep CNN models in the area of scene identification. Pretrained CNN models automatically extract the features from the dataset, and these features are millions in number. These features are then given as input to the classifier layer of the CNN model. Over the recent years, the feature selection and ensembling techniques have significantly improved in performing the scene recognition task. Much research has taken place exploring optimization algorithms, feature selection, and feature aggregation techniques to improve the classification accuracy and ensure robustness across diverse datasets.

Among optimization algorithms, such as particle swarm optimization (PSO) [13] is a metaheuristic optimization algorithm proposed by Kennedy and Eberhart in 1995. This algorithm mainly aims at finding the optimal values for specific parameters to fulfill the design requirements with the lowest possible complexity and cost. This algorithm mainly concentrates on solving the optimal combination of feature compensation coefficients for multiple computation factors so that the recognition efficiency of the system can be improved, but the limitation of this algorithm is that it can be trapped in local minima if the diversity of the data is reduced, requiring tuning of parameters like inertia weight and learning coefficients. Genetic algorithm (GA) is an optimization algorithm that was introduced by Holland in the 1970s [14] and has since been widely applied in the domains of engineering, artificial intelligence, and image classification tasks. GA searches for the number of layers, learning rate, batch size, and activation functions for the optimal combination to maximize classification performance. It can handle complex and large datasets. It automatically selects the high-performing network architecture. Its search space includes both network structure configuration and hyperparameters. But these algorithms are computationally expensive for large populations or deep networks and have slower convergence compared to gradient-based methods. Ant colony optimization (ACO) [15] is an optimization algorithm introduced by Dorigo in the early 1990s [16]. ACO was originally developed for combinatorial problems such as the traveling salesman problem; it has been successfully adapted for optimization tasks in image classification. ACO has been used to optimize various neural network platforms, such as the number of hidden layers and learning rates, and it is used to select the most relevant subset of feature data from high-dimensional complex data, which helps in reducing the computational complexity and increases the classification accuracy and interpretable classification models. But these are computationally intensive algorithms and may need integration with local search techniques to improve efficiency, and careful tuning is required to avoid premature convergence.

The research has been conducted based on multi-modal feature selection, which refers to selecting features from heterogeneous data modalities such as images, depth maps, text, and audio. In contrast, multi-model feature selection focuses on leveraging multiple deep learning architectures to extract and combine features from the same data modality. Rather than integrating data from different sources, multi-model techniques emphasize the diversity of representations learned by models such as ResNet, AlexNet, and DenseNet. This research work is inspired by the multi-modal techniques concept of combining diverse information sources. Multi-model feature selection is particularly suitable for scenarios where a single modality is available, but richer feature representations are desired.

3. METHOD

The proposed flow diagram is depicted in Figure 1. This research work has designed our model using multiple pre-trained models to efficiently recognize the different classes of scene images with high

accuracy. The detailed proposed method for classifying scene images is shown in Figure 2. This proposed method introduced an approach to perform scene recognition by combining the CLAHE and Retinex illumination technique [17], with the CD-WCO algorithm for selecting diversified features. This unique combination method of the CLAHE and Retinex illumination technique with the CD-WCO algorithm enhances classification accuracy and addresses various challenges, such as redundancy, overfitting, and the removal of irrelevant features. Feature extraction plays a prominent role in the image classification task. In the traditional method, a single deep learning model was used to extract the features from the images, which may be limited in its capacity to capture diverse and complementary information.

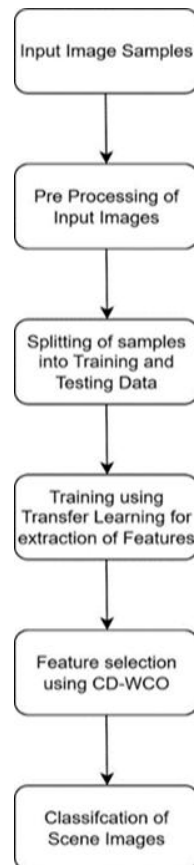


Figure 1. General flow diagram of the proposed methodology

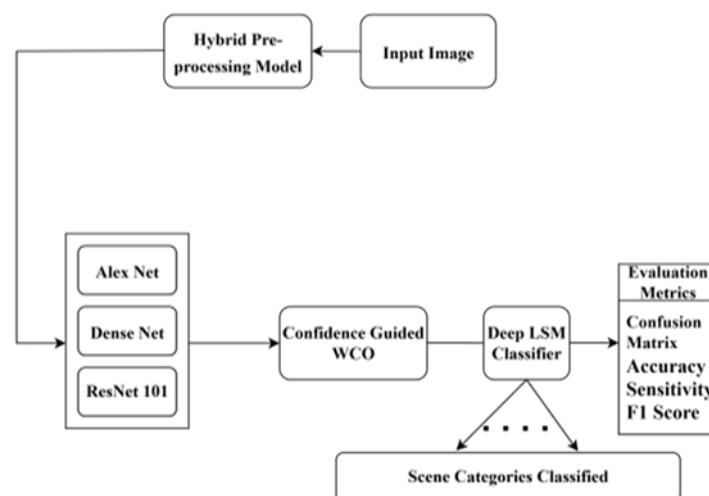


Figure 2. General block diagram of our proposed methodology with the feature selection technique

In this proposed work of multi-model feature extraction have considered ensembling the features extracted from multiple models and selecting the superior features. The feature extraction summary table for the multiple models used in this proposed work is shown in Table 1. Table 1 represents the features extracted from three widely used deep learning architectures when applied to the NYU indoor scene dataset, consisting of 1,920 images. Features are collected from the last fully connected layer (FC8 in AlexNet, FC1000 in ResNet101) or global average pooling layer (avg_pool in DenseNet201). These layers provide compact, high-level semantic representations of images. Table 1 represents the quantitative evidence that extracted features can become large and redundant. Therefore, applying a feature selection method is very crucial.

The concatenation approach allows the model to preserve both the low-level discriminative power of AlexNet and the high-level semantic abstractions from ResNet and DenseNet. The ensemble acts as a consensus mechanism: the strengths of one model compensate for the weaknesses of another, resulting in a more balanced and robust scene classification framework. The ResNet model uses skip connections for the flow of gradients and enables the learning of deeper features without degradation. AlexNet is one of the pioneering CNN architectures, which facilitates identifying prominent low- and mid-level features such as edges, shapes, and textures. DenseNet enhances feature reuse, and dense connectivity helps in learning detailed high-level representations that encode spatial and semantic relationships within the scene. This research work proposed a feature selection technique based on the CD-WCO technique, which selects the superior features and then feeds the selected features to the next stage of classification. Some of the contributions are summarized as follows:

- Implementation of a hybrid preprocessing technique, which includes CLAHE and Retinex illumination corrections, and integrating these pre-processing techniques with the CD-WCO reduces noise amplification and enhances the visibility of fine details and textures.
- Feature extraction using multiple models enables diversified learning instead of the extraction of features using a single neural network model.
- DLSM [18], used as a classifier in the final stage, provides state-based classification decisions, leading to improved accuracy and stability.

The steps involved in our proposed work for feature selection are summarized as follows:

- i) The algorithm input is the features extracted from the multiple models.
- ii) The confidence score for each solution is calculated.
- iii) For each of the feature subsets obtained, the fitness value is calculated using the predefined fitness function.
- iv) The features with a higher value of confidence are selected as optimal attributes.
- v) Features with low confidence values are discarded.
- vi) The algorithm output is the indexed feature values used for classification purposes.

Table 1. Feature dimensionality of pre-trained models on the NYU indoor scene dataset

Model	Size of the input image	Feature extraction layer	Feature vector
AlexNet	$227 \times 227 \times 3$	FC8	1920×4096
ResNet101	$224 \times 224 \times 3$	FC1000	1920×2048
DenseNet 201	$224 \times 224 \times 3$	Avg pool layer	1920×1920

3.1. Dataset and pre-processing

In this work, experimentation is performed on the NYU [19] and MIT Indoor-15 class subset [3] datasets. This work used a hybrid pre-processing technique, which includes CLAHE and Retinex illumination corrections, which improve the contrast of the images and normalize the lighting variations in the images. CLAHE makes the small details more visible in the images. The fusion of both gives illumination-normalized and detail-enhanced images, which provide richer and cleaner features for deep learning models. CLAHE and Retinex are applied sequentially. CLAHE ensures local contrast enhancement, while Retinex corrects global illumination inconsistencies. The combination produces images with balanced brightness, improved contrast, and reduced lighting artifacts. This hybrid preprocessing pipeline enables models to focus on meaningful structural and textural details, which is crucial for scene classification tasks where lighting variations are common. The joint application of Retinex and CLAHE, followed by CD-WCO, ensures both global illumination normalization and localized contrast amplification. Retinex models the human visual system to maintain natural reflectance under varying illumination, while CLAHE adaptively enhances micro-contrast across spatially distinct regions. Subsequently, CD-WCO performs channel-wise dual-weighted fusion to balance chromatic and structural information.

3.2. Confidence-based world cup optimization-based feature selection

Feature selection is a very important and crucial step in deep learning. Let $F = \{f_1, f_2, f_3, \dots, f_n\}$ denote the feature set extracted from multiple pre-trained models (ResNet, DenseNet, and AlexNet). The optimization process seeks an optimal subset $F^* \subseteq F$ that maximizes classification accuracy while minimizing redundancy. To improve the classification efficiency for the entire set of extracted features obtained from the pre-trained model, the passes through the confidence-based weighing mechanism. For each of the features f_i , a confidence-based calculated, and it is represented as a presence $C(f_i)$. The confidence score is calculated by (1).

$$C(f_i) = \frac{1}{M} \sum_{j=1}^M I(y_j = \hat{y}_j | f_i) \quad (1)$$

Where M is the total number of validation samples; y_j is the ground truth label of sample j ; $\hat{y}_j$ is the predicted label using feature f_i ; I is an indicator function that equals 1 if the condition is true and 0 otherwise; and $C(f_i) \in [0,1]$ represents the reliability of feature f_i . This confidence score calculation quantifies the reliability of each feature to perform classification.

3.3. Fitness function in the proposed methodology

The fitness function is the key criterion that guides the search towards optimal solutions. In the case of our proposed methodology of feature selection, the fitness function evaluates the quality of a candidate feature subset by balancing classification accuracy, confidence, reliability, and feature subset compactness. $F_k \subseteq F$ be a candidate feature subset, where F is the complete feature set extracted from the ensemble of pre-trained models. The fitness of F_k is given as (2).

$$\tau(F_k) = \alpha Acc(F_k) + \beta \frac{1}{|F_k|} \sum_{f_i \in F_k} C(f_i) - \gamma \cdot \frac{|F_k|}{|F|} \quad (2)$$

Where $Acc(F_k)$ is classification accuracy obtained using subset F_k ; $C(f_i)$ is confidence score of features f_i ; $|F_k|$ is number of selected features in subset F_k ; $|F|$ is total number of available features; and α, β, γ is control parameters that balance the contributions of accuracy, confidence, and dimensional reduction.

The fitness function in confidence-based WCO is a multi-objective evaluation measure that ensures the final selected feature subset is not only accurate but also robust and efficient. This approach is well-suited for high-dimensional feature spaces generated by deep ensembles like ResNet, DenseNet, and AlexNet. By applying confidence-based WCO, this work obtains an optimal subset F^* that contains only the most informative, reliable, and compact features. This reduces computational overhead, enhances model interpretability, and improves the generalization ability of the classifier across diverse scene categories. Mid-level fusion was chosen as it offers a balanced trade-off between representational richness and computational efficiency. Unlike early fusion, which operates directly on raw inputs and may introduce redundancy, and decision-level fusion, which combines independent classifier outputs and may lose inter-feature correlations, mid-level fusion preserves discriminative semantic relationships while enabling effective dimensionality reduction through optimization.

3.4. Deep liquid state machine classifier

At the final stage of the methodology, the DLSM [20], [21] is used as the classifier that maps the extracted features into the corresponding scene categories. Unlike conventional classifiers (e.g., softmax layers or support vector machines (SVMs)), the DLSM leverages the dynamic temporal encoding properties of liquid states combined with the hierarchical learning ability of deep networks, making it particularly effective for complex scene data [21]. The proposed DLSM architecture is shown in Figure 3. LSM consists of three special layers: an input layer, a reservoir or liquid layer, and a memory-less readout circuit. The reservoir is considered the generic preprocessor that consists of numerous leaky integrate-and-fire (LIF) neurons. Readout neurons are also called processors, and they produce the final output from the LSM.

The hyperparameters used for the DLSM classifier are as shown below. The spectral radius of the reservoir was chosen to be less than one in order to maintain stable reservoir dynamics and satisfy the echo-state property. The leak rate was selected to control the rate at which past information influences the current neuron state, thereby allowing the model to retain useful temporal information without causing instability. To reduce computational complexity, sparse reservoir connectivity was adopted, which limits the number of active connections while preserving sufficient representational capability. In addition, the SWTA mechanism was employed to suppress weak neuron activations and emphasize dominant responses, leading to improved class separability during the readout stage as shown in Table 2.

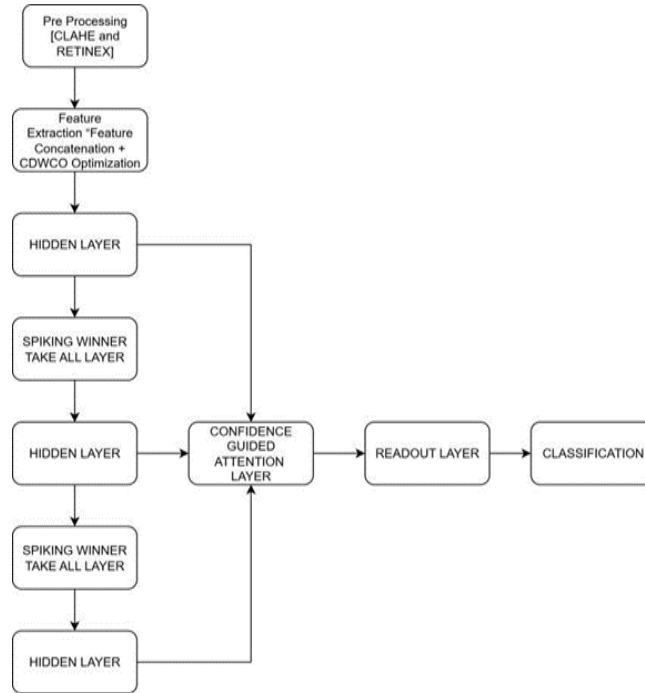


Figure 3. Proposed DLSSM architecture with confidence-guided attention and SWTA layers for scene recognition

Table 2. Hyperparameter settings of the SWTA reservoir and readout layers

Parameter	Description	Value
Reservoir size	Number of spiking neurons	500
Input scaling	Scaling of input feature weights	0.5
Spectral radius	Controls reservoir stability	0.9
Leak rate	Neuron state update rate	0.3
Sparsity	Percentage of non-zero reservoir connections	10%
Activation function	Reservoir neuron activation	tanh
Readout layer	Linear regression/softmax	Linear
Learning rate	Readout learning rate	0.001
Training epochs	Number of training iterations	100
Regularization	L2 regularization coefficient	1e-4
SWTA threshold	State selection threshold	0.8

DLSMs are typically used for temporal data, and static images can be transformed into pseudo-temporal spike sequences through rate coding. Introducing a SWTA layer ensures that only the most active neurons contribute to the liquid state. Each image feature (from the CNN ensemble) is first converted to spike rates. Given a feature vector $F = [f_1, f_2, \dots, f_N]$, as in (3).

$$S_i(t) = \{1, \text{ if } t \leq r_i T_{max} \text{ } 0, \text{ otherwise}\} \quad (3)$$

Where $S_i(t)$ is spike train for neuron; $r_i = \frac{f_i - f_{min}}{f_{max} - f_{min}}$ is normalized firing rate; and T_{max} is simulation window (e.g., 100 ms). The SWTA layer selects only the most dominant spikes per time step. For N input neurons, the membrane potential is (4).

$$V_i(t) = \sum_j W_{ij} S_j(t) - \theta \quad (4)$$

At each time step, only the neuron with the maximum potential fires, as shown in (5).

$$O_k(t) = \{1, \text{ if } k = \arg \max_i V_i(t) \text{ } 0, \text{ otherwise}\} \quad (5)$$

This enforces competition among neurons; only the winner emits a spike, improving feature selectivity and sparsity. The SWTA—filtered spikes enter the liquid layer of $N_r = 1000$ LIF neurons, with

random sparse connectivity (10%). As shown in Figure 3, the optimized features are processed through spiking reservoir layers with SWTA competition, followed by a confidence-guided attention mechanism and a readout layer for final scene classification across 20 categories. The reservoir dynamics transform static features into high-dimensional temporal representations, which are classified via a softmax readout. This integration leverages both the diversity of CNN features and the temporal richness of spiking reservoirs, achieving a final accuracy of 96.5%. 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 Euclidean distance, $D(p, q)$, between them as shown (6).

$$P(p, q) = C * e(-D(p, q) / \tau) 2 \quad (6)$$

3.5. Dataset

This proposed work is carried out using the NYU depth dataset V2. The dataset consists of 20 different scene categories and over 1,000 object classes. For this work have worked on a subset of 1,920 indoor scene images with a total size of 79.5 MB. These images exhibit variations in spatial layout, orientation, and scene composition and were systematically sorted and reorganized into 20 distinct scene classes. The dataset consists of indoor scenes such as basement, bathroom, bedroom, book shop, cafe, classroom, conference room, dinette, dining room, exercise room, furniture shop, home office, home storage, kitchen, computer lab, living area, play room, reception, student lounge, and study room, as shown in Figure 4. This proposed work has 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.

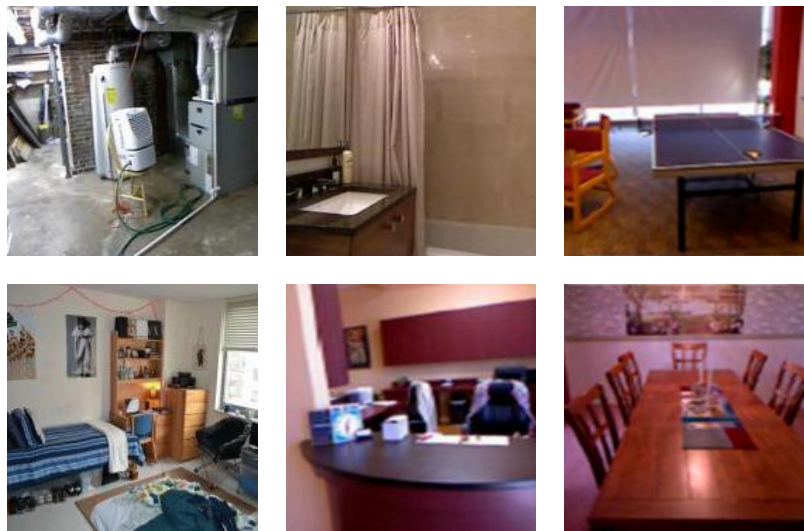


Figure 4. Sample scene images of various classes such as basement, bathroom, play room, bedroom, reception, and dining room in the NYU dataset

4. RESULTS AND DISCUSSIONS

This research work has evaluated the performance of the proposed methodology on the NYU dataset, which includes different classes of scene images. To enhance the quality of the input scene images, this work has applied the CLAHE and Retinex illumination corrections as a preprocessing technique to enhance the quality of the images. It reduces the background noise, as shown in Figure 5, where Figure 5(a) shows the sample original input image, Figure 5(b) shows the CLAHE applied, and Figure 5(c) shows the Retinex applied. The dataset consists of different scene classes such as bedroom, dining room, bathroom, and study room. CLAHE aims to improve local contrast, especially in areas that are too dark or too bright, whereas focus on improving the dynamic range and color constancy of an image, often making details more visible in both bright and dark regions.

The integration of CLAHE and Retinex preprocessing significantly enhances the scene recognition performance across all backbone networks. The proposed DLSSM with SWTA classifier achieves 96.5% accuracy, reflecting a 2.7% improvement as shown in Table 3. This demonstrates that illumination normalization enhances feature consistency, benefiting both CNN feature extraction and spiking neuron

stability during classification. Table 3 shows a comparative evaluation of classification accuracy obtained by different deep learning models with and without the proposed pre-processing strategy on the NYU indoor scene dataset. Table 3 clearly demonstrates the impact of illumination-aware pre-processing on both individual CNN models and ensemble-based approaches.



Figure 5. The sample input image and the images after application of (a) original input, (b) CLAHE, and (c) Retinex

Table 3. Performance comparison before and after preprocessing for the NYU and MIT Indoor scene-15 dataset

Model	Without pre processing (%) [NYU dataset]	With CLAHE + Retinex (%) [NYU dataset]	Without pre processing (%) [MIT indoor 15 dataset]	With CLAHE + Retinex (%) [MIT indoor 15 dataset]
AlexNet	89.80	92.70	86.41	89.61
ResNet	92.52	95.31	89.22	92.12
DenseNet	90.60	93.11	88.00	91.04
Multi-model ensemble (concatenation)	94.12	96.22	91.81	94.60
Proposed DLSSM with SWTA	93.80	96.50	92.40	95.20

Without preprocessing, AlexNet achieves an accuracy of 89.8%, indicating its sensitivity to illumination variations and background clutter due to its relatively shallow architecture. ResNet and DenseNet show improved baseline performance (92.5% and 90.6%, respectively) owing to deeper architectures and improved feature reuse; however, they still exhibit performance degradation under uneven lighting conditions. After applying the hybrid CLAHE + Retinex preprocessing, all models show a consistent and notable improvement in accuracy. AlexNet improves to 92.7%, while ResNet and DenseNet achieve 95.3% and 93.1%, respectively, confirming that pre-processing enhances feature quality by improving contrast and illumination consistency. The multi-model ensemble using feature concatenation further improves performance to 96.2%, demonstrating the benefit of combining complementary representations from multiple architectures.

However, the proposed DLSSM with achieves the highest accuracy of 96.5%, outperforming both individual models and the conventional ensemble approach. This improvement highlights the effectiveness of the DLSSM's memory-based learning and the SWTA mechanism in leveraging enhanced feature representations more efficiently than static fusion strategies. Overall, the results in Table 4 validate that the proposed pre-processing not only boosts individual CNN performance but also synergistically enhances the proposed DLSSM–SWTA framework, leading to superior and more stable indoor scene classification performance. From the results in Table 4, it is evident that CD-WCO consistently reduces the number of selected features across all deep learning models. CD-WCO eliminates nearly one-fourth of the redundant features. This reduction is attributed to the incorporation of confidence scores before the optimization process, which enables the algorithm to discard features with low reliability. CD-WCO improves computational efficiency by lowering storage and training costs, while simultaneously enhancing the robustness of the final classification stage. The retained features are more discriminative, which improves classifier training speed, reduces memory usage, and enhances generalization. The bar graph representing the feature reduction is shown in Figure 6.

As depicted in Figure 6, across all three models, the WCO algorithm reduces the number of features extracted from the models. Feature dimensionality gets reduced. The CD-WCO further improves the performance by consistently reducing the number of features and selecting the diversified and discriminative features. This represents that the confidence-driven selection enhances the efficiency of feature extraction, leading to more compact representations that are computationally less expensive and potentially more robust for classification tasks. The proposed method is evaluated on datasets consisting of 1,900 images and across

20 different categories of scene images in the dataset. The dataset has been split into 80% training and 20% testing results in 1,536 training images and 384 testing images. This work has employed the Adam optimizer with a learning rate of 0.01, and the training process was for 500 iterations and for a mini batch size of 128. The plot of accuracy and loss curves is as shown in Figures 7 and 8. Figure 7 shows that the model is learning better to obtain an accuracy of 96.5%. In this simulation, the validation accuracy also increases and stabilizes at a slightly lower percentage around 94%. Figure 8 shows that the model's loss is reduced in the successive epochs.

Table 4. Reduction in feature dimensionality through confidence-driven WCO compared with standard WCO across deep learning models on the NYU scene dataset

Model	Feature extraction layer	Features extracted	Features extracted using WCO	Features extracted using CD-WCO	Reduction percentage (%)
AlexNet	FC8	23,572,480	11,648,120	8,736,090	25.00
ResNet101	FC1000	11,786,240	5,881,610	4,411,207	25.00
DenseNet201	avg pool	11,049,600	5,588,105	4,191,079	25.00

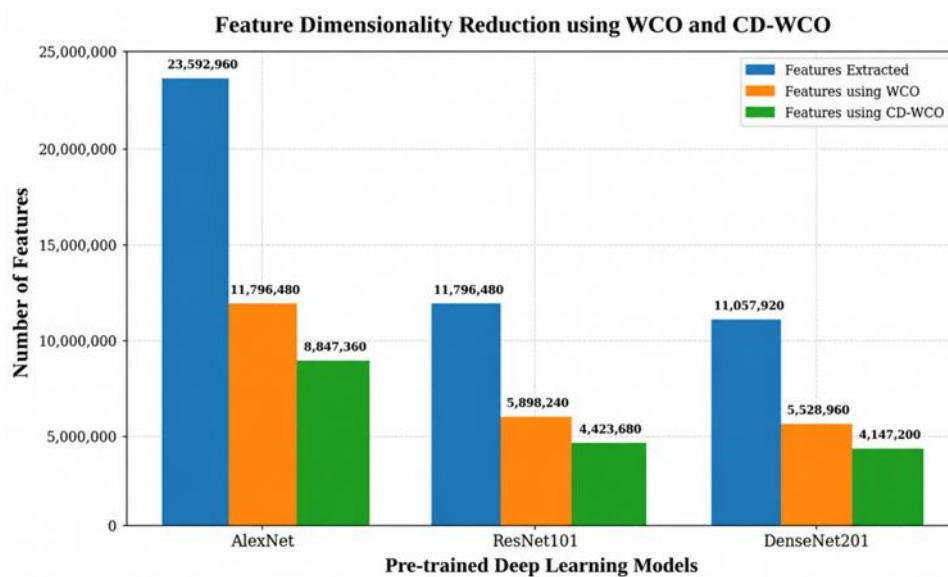


Figure 6. Feature dimensionality for pre-trained models compared with those reduced using WCO and CD-WCO

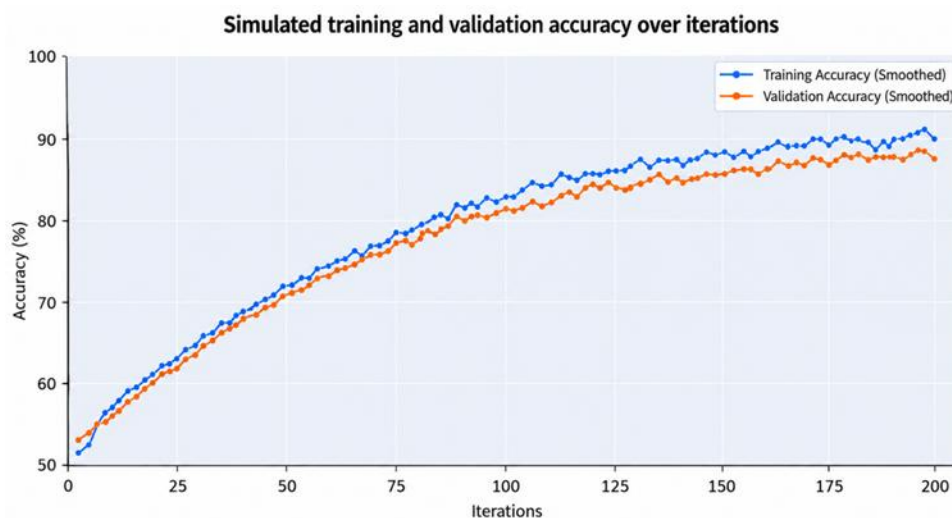


Figure 7. Plot of accuracy curves for the scene image NYU dataset

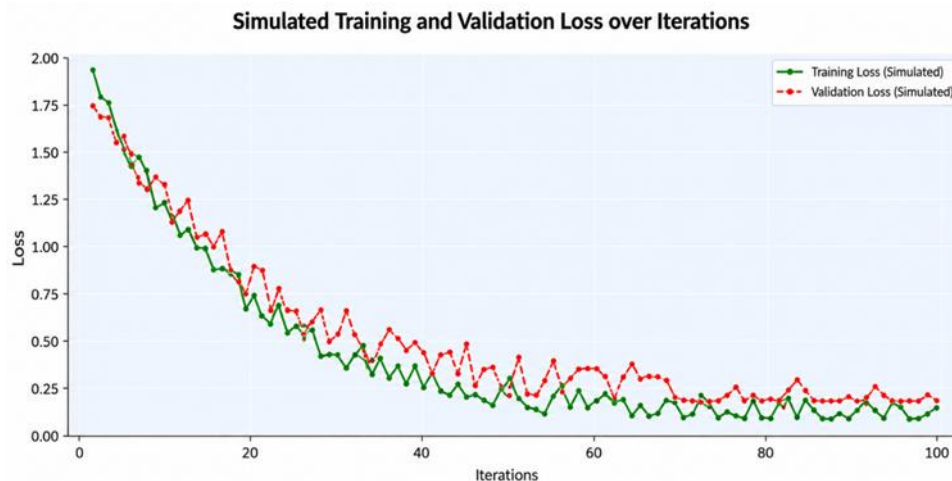


Figure 8. Plot of loss curves for the scene image NYU dataset

According to Figure 8, the green line represents the loss of the model on training data. The representation shows that the loss is decreasing as the number of iterations increases, and shows that the model is learning properly to perform the classification. All the experimentation was carried out by using the Python language and a powerful image processing environment. This work used a system equipped with an Intel Core i7 processor with an operating speed of a speed of 2.30 GHz, an NVIDIA RTX GPU, and 16 GB of system memory. Table 5 indicates the experimentation conducted on the SCENE-15 dataset. Table 5 shows the progressive reduction of feature dimensionality at different stages of processing. Features extracted from multiple layers of neural networks, such as AlexNet, ResNet101, and DenseNet201 have dimensions of 4,096, 2,048, and 1,920, respectively. The features extracted from multiple layers are combined to form a vector of 8,064 dimensions.

The final vector of 8,064 dimensions is further reduced to a 4,032-feature vector, but the vector still contains redundant features. To solve this problem, the confidence-guided weighted combination (CD-WCO) algorithm is applied to reduce the feature dimension from 4,032 to 2,150, achieving a 46.7% reduction. At the final stage, the DLSSM with the SWTA mechanism further refines the feature space by activating dominant memory neurons, resulting in a compact representation of 1,280 dimensions.

Table 5. Feature dimensionality and reduction percentage on the MIT Indoor scene-15 dataset

Model/stage	Original feature dimension	Reduced feature dimension	Reduction (%)
AlexNet (FC7)	4,096	2,048	50.00
ResNet101 (Global avg pool)	2,048	1,024	50.00
DenseNet201 (Global avg pool)	1,920	960	50.00
Multi-model feature concatenation	8,064	4,032	50.00
After CD-WCO optimization	4,032	2,150	46.70
Final features (DLSSM + SWTA)	2,150	1,280	40.50
Overall reduction (initial → final)	8,064	1,280	84.11

Figure 9 represents the training and validation accuracy (Figure 9(a)) and loss (Figure 9(b)) plots for the MIT Indoor scene-15 dataset. The accuracy curve demonstrates steady improvement in accuracy and gradually attains a stable value of 95.32% at 30 epochs. The minimum difference between training and validation accuracy indicates good generalization performance. Correspondingly, the loss curves demonstrate consistent decay, confirming stable optimization and effective learning behavior of the proposed DLSSM-SWTA framework. Figure 10 confusion matrix obtained for the MIT Indoor scene-15 dataset provides a detailed class-wise evaluation of the proposed scene recognition framework. A strong diagonal dominance is observed across most classes, demonstrating that the proposed model effectively discriminates between different scene types.

In Figure 11, the confusion matrix obtained for the NYU dataset using the proposed methodology is shown. In the output, as depicted, the value of diagonal elements is high, indicating that the model is learning and performing well and classifying most of the scene samples correctly. A smaller number, other than the diagonal elements, shows the mistakes made by the model. The plot also includes the overall accuracy (96.58%), which is the proportion of correct predictions (the sum of the diagonal values) to the total number of data points in the test set. In this work has proposed a framework that integrates preprocessing techniques

such as CLAHE and Retinex, transfer learning-based feature extraction, where we have ensembled features from models, such as AlexNet, ResNet, DenseNet, and later using confidence-based feature selection approach to select the desired features, which in turn reduces the number of computations and network complexity also reduces and accuracy has been improved compared to other models.

Table 6 shows the comparison of this proposed methodology with the existing techniques. This proposed methodology shows that the proposed method outperforms other existing methodologies. Table 7 validates the effectiveness, stability, and generalization capability of the proposed framework for large-scale indoor scene classification. The performance metrics indicated in Table 7 show the contribution of each block in the proposed methodology.

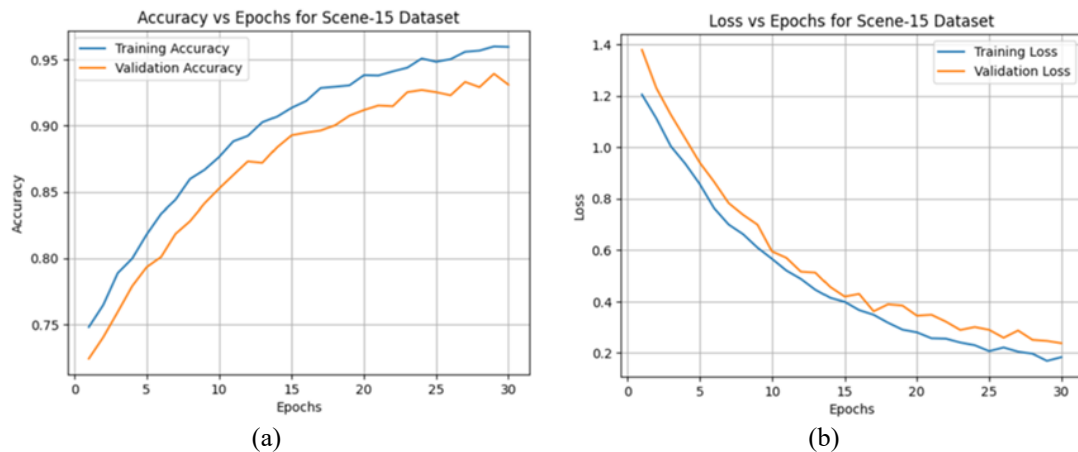


Figure 9. Training and validation curves of the proposed DLSM-SWTA framework on the scene-15 dataset: (a) accuracy and (b) loss

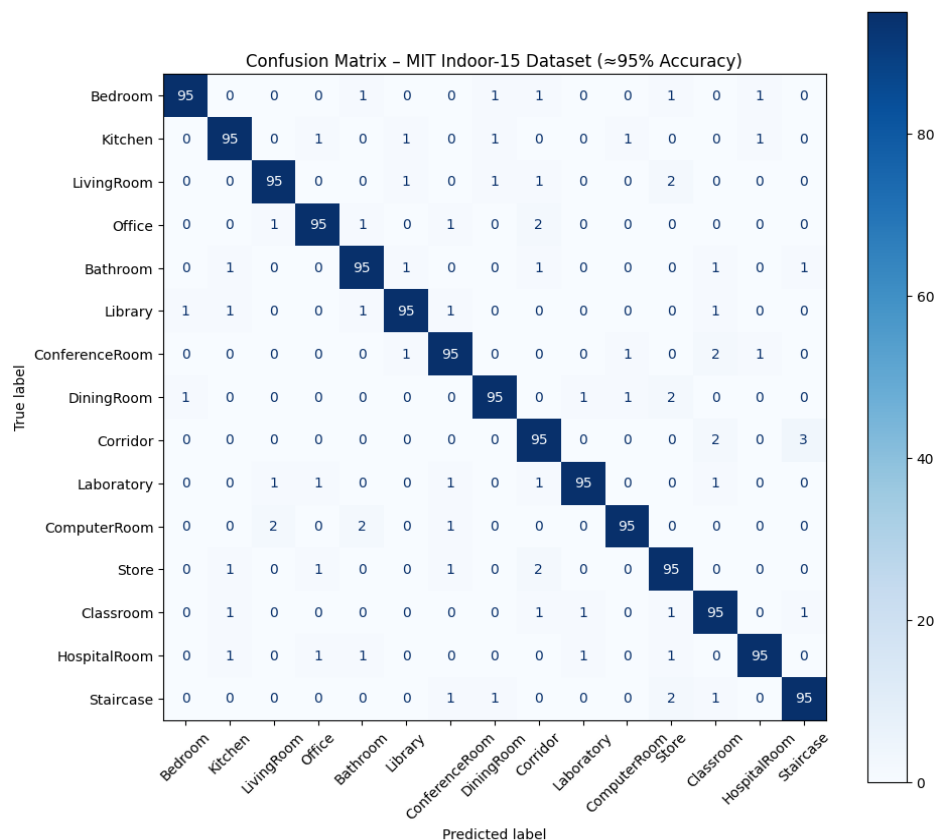


Figure 10. Confusion matrix of the proposed DLSM-SWTA framework on the MIT Indoor scene-15 dataset

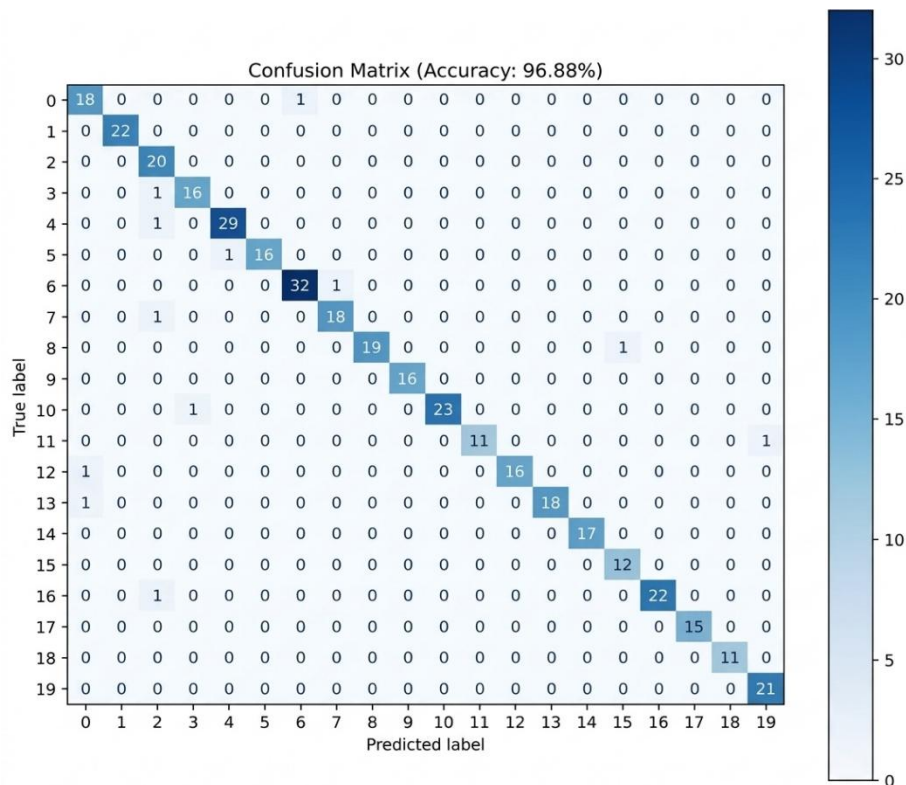


Figure 11. Confusion matrix obtained for the NYU dataset

Table 6. Accuracy comparison of the proposed model with existing works

No	Author	Methodology	Accuracy (%)
1	Pereira <i>et al.</i> [22]	Semantic segmentation + VGG16, ResNet18-50-101, DenseNet, and MobileNetV2 + feature fusion	75.80
2	Li <i>et al.</i> [23]	MAPNet + attentive pooling	67.70
3	Tang <i>et al.</i> [24]	Multi-feature fusion + GoogLeNet	92.92
4	Afif <i>et al.</i> [25]	Scaling + EfficientNet CNN model	95.60
5	Proposed method	Preprocessing + DenseNet201, AlexNet, ResNet + confidence guided world cup optimization + DLSSM classifier	96.52

Table 7. Class-wise performance comparison of the proposed hybrid model on the NYU dataset with 20 scene classes

Metric	Living Room(0)	Bed Room(1)	Kitchen(2)	Office	Bath Room	Corridor	Dining room	Class room	Library
Precision	0.97	0.96	0.95	0.97	0.96	0.95	0.96	0.97	0.98
Recall	0.96	0.97	0.96	0.95	0.97	0.96	0.97	0.96	0.97
F1-score	0.96	0.96	0.95	0.96	0.96	0.95	0.96	0.96	0.97
Support	280	300	260	240	230	220	250	270	210

Table 8 shows the class-wise comparison of performance metrics using the proposed methodology for 9 different classes. Table 9 shows the class-wise comparison of performance metrics using the proposed methodology for warehouse and workshop categories on the NYU dataset and the weighted average and macro average of performance metrics for all 20 classes of NYU dataset. The hybrid preprocessing block improves the image quality by enhancing the local contrast and compensating for illumination variations, as consistently shown in the higher recall values across all the scene categories in the NYU dataset. Extracting the features from the multi-model blocks captures features from the various levels, which contributes to higher precision values. The CD-WCO block adaptively fuses features by assigning higher weights to more confident representations, rather than performing simple feature concatenation. The deep LSM classifier operates on the fused feature space and effectively contributes to the overall classification accuracy.

Table 8. Class-wise performance comparison of the proposed method on NYU dataset

Metric	Store	Basement	Conference room	Computer room	Gym	HallWay	Laundry room	Study room	Staircase
Precision	0.96	0.95	0.96	0.97	0.95	0.96	0.95	0.97	0.96
Recall	0.96	0.95	0.96	0.96	0.95	0.96	0.95	0.97	0.96
F1-score	0.96	0.95	0.96	0.96	0.95	0.96	0.95	0.97	0.96
Support	240	200	190	210	180	220	170	200	160

Table 9. Class-wise performance comparison of the proposed method for the warehouse and workshop categories on the NYU dataset

Metric	Warehouse(18)	Workshop(19)	Macro avg	Weighted avg
Precision	0.95	0.96	0.96	0.97
Recall	0.95	0.96	0.96	0.97
F1-score	0.95	0.96	0.96	0.97
Support	180	190	4,130	4,130
Overall-accuracy	Warehouse(18)	Workshop(19)	0.965	0.965

Figures 12 and 13 illustrate the class-wise F1-score, precision, and recall values achieved by the proposed framework across 20 indoor scene categories of the NYU dataset. The F1-score, which balances precision and recall, is used to assess the model’s robustness in handling both false positives and false negatives. As shown in the bar chart, the proposed method maintains consistently high F1-scores ranging from 0.95 to 0.97 across all scene classes. This uniform performance indicates that the model effectively discriminates between visually similar indoor environments, such as living rooms, bedrooms, and offices, without bias toward specific categories. As observed, precision and recall values are consistently high across all the scene categories, indicating reliable classification performance and minimal bias toward specific classes. The close alignment between the precision and recall bars for most scene types suggests that the model maintains a balanced trade-off between correctly identifying scene instances and avoiding false classifications.

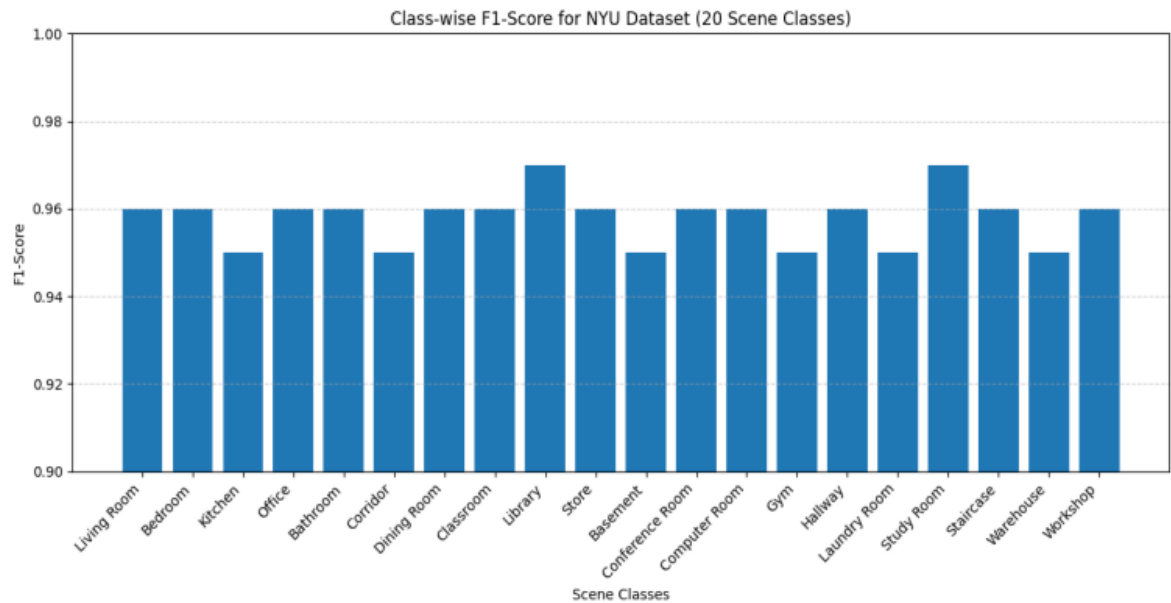


Figure 12. Bar plot showing the class-wise F1-score for the NYU dataset using the proposed methodology

Table 10 DLSTM classifier performance is compared against widely used conventional and deep learning-based classifiers under identical experimental settings. All methods use the same feature representations extracted from AlexNet, ResNet, and DenseNet to ensure an unbiased comparison. The comparative results presented demonstrate that the classical machine learning classifiers, such as SVM, k-nearest neighbor (KNN), and random forest (RF), show limited performance due to their reliance on static

decision boundaries and reduced capability to model complex non-linear feature interactions. Deep learning classifiers such as softmax-based CNN heads and long short-term memory (LSTM) classifiers achieve improved performance; however, they are less effective in capturing the dynamic temporal evolution of deep feature representations. The proposed method of DLSTM leverages reservoir-based dynamics and temporal state transitions, enabling richer feature encoding and enhanced class separability. DLSTM achieves the maximum accuracy of 96.5%, with superior precision, F1-score, and recall, validating its suitability for complex scene recognition tasks.

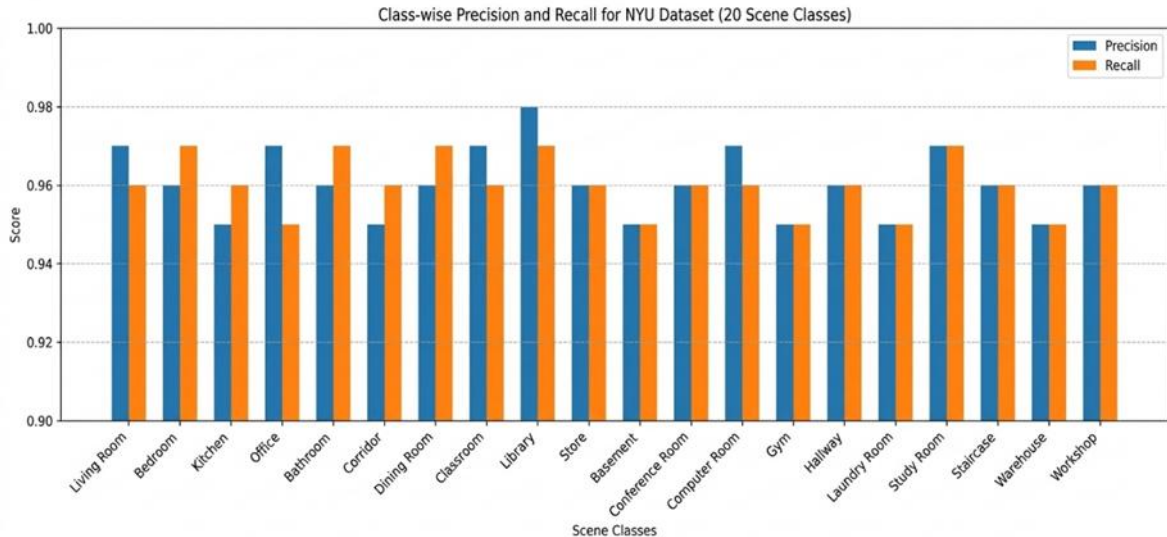


Figure 13. Plot of precision and recall values for each indoor scene class using the proposed method on the NYU dataset

Table 10. Performance comparison of the proposed DLSTM with SWTA against existing classifiers

Classifier	Accuracy (%)	Precision	Recall	F1-score
SVM [26]	90.8	0.91	0.90	0.90
KNN [27]	88.6	0.89	0.88	0.88
RF	91.9	0.92	0.92	0.92
Softmax (CNN classifier)	93.4	0.93	0.93	0.93
LSTM classifier [28]	94.1	0.94	0.94	0.94
Proposed DLSTM with SWTA	96.5	0.96	0.97	0.96

5. CONCLUSION

In this paper proposed a model that performs scene recognition using a transfer learning methodology that ensembles the features extracted from the pretrained models. These extracted features will be passed to the confidence-based feature selection block to select the relevant features based on the confidence score obtained during the processing stage and optimized using the WCO technique. Only the selected features are processed further for the classification process, and classification is performed using the deep LSM classifier. In future work will concentrate on working with different datasets and combining different modalities in order to increase the applicability of this proposed work. This work has a wide range of applications in helping visually impaired and senior citizens with indoor navigation

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Sumathi	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	
Krishnamurthy														
Shetty Pramod Kumar		✓				✓		✓		✓		✓		
Halasur Rudrappa	✓		✓	✓		✓	✓		✓		✓	✓	✓	
Mahadevaswamy														
Ujwala Bagepalli			✓		✓		✓			✓		✓		
Satishbabu														

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P : Project administration

Fu : Funding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author, [SK], upon reasonable request.

REFERENCES

- [1] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: spatial pyramid matching for recognizing natural scene categories," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006, pp. 2169–2178, doi: 10.1109/CVPR.2006.68.
- [2] A. Oliva and A. Torralba, "Modeling the shape of the scene: a holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001, doi: 10.1023/A:1011139631724.
- [3] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 413–420, doi: 10.1109/CVPR.2009.5206537.
- [4] H. Wu, G. Tian, Y. Li, F. Zhou, and P. Duan, "Spatial semantic hybrid map building and application of mobile service robot," *Robotics and Autonomous Systems*, vol. 62, no. 6, pp. 923–941, Jun. 2014, doi: 10.1016/j.robot.2013.01.001.
- [5] L. Xie, F. Lee, L. Liu, K. Kotani, and Q. Chen, "Scene recognition: a comprehensive survey," *Pattern Recognition*, vol. 102, 2020, doi: 10.1016/j.patcog.2020.107205.
- [6] K. Alpan, B. Arman, and K. Dimililer, "Effect of contrast limited adaptive histogram equalization (CLAHE) on breast cancer detection using residual network (ResNet)," in *2025 1st International Conference on Computational Intelligence Approaches and Applications (ICCIAA)*, 2025, pp. 1–5, doi: 10.1109/ICCIAA65327.2025.11013776.
- [7] S. V.-Flores, B. Ermentrout, and M. Bodner, "Modeling neuropathologies as disruption of normal sequence generation in working memory networks," *Neural Networks*, vol. 27, pp. 21–31, Mar. 2012, doi: 10.1016/j.neunet.2011.09.007.
- [8] Y. Zhou, C. Shi, B. Lai, and G. Jimenez, "Contrast enhancement of medical images using a new version of the world cup optimization algorithm," *Quantitative Imaging in Medicine and Surgery*, vol. 9, no. 9, pp. 1528–1547, Sep. 2019, doi: 10.21037/qims.2019.08.19.
- [9] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004, doi: 10.1023/B:VISI.0000029664.99615.94.
- [10] I. T. Jolliffe, *Principal component analysis*, 2nd ed. Springer, 2002, doi: 10.1007/b98835.
- [11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015, pp. 1–14.
- [12] C. Szegedy et al., "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015, pp. 1–9, doi: 10.1109/CVPR.2015.7298594.
- [13] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of ICNN'95 - International Conference on Neural Networks*, 1995, vol. 4, pp. 1942–1948, doi: 10.1109/ICNN.1995.488968.
- [14] J. H. Holland, *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. Michigan, USA: University of Michigan Press, 1975.
- [15] Y. A. Mukhlif, A. A. Al-Shammari, H. A. Al-Mashhadani, M. A. Al-Shammari, and A. A. Al-Dabbagh, "Ant colony and whale optimization algorithms aided by neural networks for optimum skin lesion diagnosis: a thorough review," *Mathematics*, vol. 12, no. 7, 2024, doi: 10.3390/math12071049.
- [16] M. Dorigo, "Optimization, learning and natural algorithms." Ph.D. Thesis, Politecnico di Milano, Milan, 1992.
- [17] D. J. Jobson, Z. Rahman, and G. A. Woodell, "A multiscale retinex for bridging the gap between color images and the human observation of scenes," *IEEE Transactions on Image Processing*, vol. 6, no. 7, pp. 965–976, 1997, doi: 10.1109/83.597272.
- [18] D. Verstraeten, B. Schrauwen, M. D'Haene, and J. Van Campenhout, "An experimental unification of reservoir computing methods," *Neural Networks*, vol. 20, no. 3, pp. 391–403, 2007, doi: 10.1016/j.neunet.2007.04.003.
- [19] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGB-D images," in *European Conference on Computer Vision (ECCV)*, 2012, pp. 746–760, doi: 10.1007/978-3-642-33715-4_54.
- [20] N. Soares and D. Kudithipudi, "Deep liquid state machines with neural plasticity for video activity recognition," *Frontiers in Neuroscience*, vol. 13, p. 686, 2019, doi: 10.3389/fnins.2019.00686.
- [21] R. Surendran, I. Chihi, J. Anitha, and D. J. Hemanth, "Indoor scene recognition: an attention-based approach using feature selection-based transfer learning and deep liquid state machine," *Algorithms*, vol. 16, no. 9, 2023, doi: 10.3390/a16090430.

- [22] R. Pereira, T. Barros, L. Garrote, A. Lopes, and U. J. Nunes, "A deep learning-based global and segmentation-based semantic feature fusion approach for indoor scene classification," *Pattern Recognition Letters*, vol. 179, pp. 24–30, 2024, doi: 10.1016/j.patrec.2024.01.022.
- [23] Y. Li, Z. Zhang, Y. Cheng, L. Wang, and T. Tan, "MAPNet: multi-modal attentive pooling network for RGB-D indoor scene classification," *Pattern Recognition*, vol. 90, pp. 436–449, 2019, doi: 10.1016/j.patcog.2019.02.005.
- [24] P. Tang, H. Wang, and S. Kwong, "G-MS2F: GoogLeNet based multi-stage feature fusion of deep CNN for scene recognition," *Neurocomputing*, vol. 225, pp. 188–197, 2017, doi: 10.1016/j.neucom.2016.11.023.
- [25] M. Afif, R. Ayachi, Y. Said, and M. Atri, "Deep learning based application for indoor scene recognition," *Neural Processing Letters*, vol. 51, no. 3, pp. 2827–2837, 2020, doi: 10.1007/s11063-020-10231-w.
- [26] S. Rezvani, F. Pourpanah, C. P. Lim, and Q. M. J. Wu, "Methods for class-imbalanced learning with support vector machines: a review and an empirical evaluation," *Soft Computing*, vol. 28, no. 20, pp. 11873–11894, Oct. 2024, doi: 10.1007/s00500-024-09931-5.
- [27] R. K. Halder, M. N. Uddin, M. A. Uddin, S. Aryal, and A. Khraisat, "Enhancing k-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications," *Journal of Big Data*, vol. 11, no. 1, Aug. 2024, doi: 10.1186/s40537-024-00973-y.
- [28] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

BIOGRAPHIES OF AUTHORS



Sumathi Krishnamurthy     received her B.E. degree in Electronics and Communication Engineering from Visvesvaraya Technological University Belagavi, M.Tech. in Digital Electronics and Communication Systems from Jawaharlal Nehru New College of Engineering, Shivamogga under Visvesvaraya Technological University Belgaum and presently working as an assistant professor in Department of Electronics and Communication Engineering of Jawaharlal Nehru New College of Engineering, Shivamogga. Her areas of interest are artificial intelligence and python programming. She can be contacted at email: sumathik@jnnce.ac.in.



Shetty Pramod Kumar     received his B.E. degree in Electronics and Communication Engineering from VEC Bellary and M.Tech. in VLSI design and embedded systems from UTL Technologies Limited, Bangalore. PhD from VTU-RRC Belgaum and working as faculty in the Department of Electronics and Communication Engineering, Jawaharlal Nehru New College of Engineering, Shivamogga, Karnataka, India. He can be contacted at email: pramod.s86@gmail.com.



Halasur Rudrappa Mahadevaswamy     is working as vice chancellor Jagadguru Sri Shivarathreeshwara University, Noida. He has received an award for contribution in 'continuous improvement process-CIP', distinguished chief guest and keynote speaker at 3D printing - the future in 2016 at the networking; workshop. Published more than 16 research papers. He co-founded a company, Applied Invention India Pvt Ltd, in 2012. He has worked as a director at Curio System Pvt Ltd. and worked as a Program Manager at Satyam, Hyderabad. Worked as a group manager at Robert Bosch Engineering and Business Solutions Ltd., Engineering. He can be contacted at email: dhrmswamy@gmail.com.



Ujwala Bagepalli Satishbabu     received her B.E. degree in Electronics and Communication Engineering from Visvesvaraya Technological University Belagavi and M.Tech. first rank with a gold medal in Digital Electronics and Communication Systems from MCE, Hassan, under VTU Belgaum, and from 2005 working as an assistant professor in the Department of Electronics and Communication Engineering of Jawaharlal Nehru New College of Engineering Shivamogga. Her areas of interest are artificial intelligence and Python programming. She can be contacted at email: ujwalaravi2004@jnnce.ac.in.