

MACHINE LEARNING APPROACHES TO NATURAL SCENES CLASSIFICATION: A COMPREHENSIVE REVIEW

Priti¹
Dinesh Singh
Savita Ahlawat

Received 24.03.2025

Revised 17.05.2025

Accepted 21.06.2025

Keywords:

Natural Scene Classification, CNN,
Deep Learning, Machine Learning,
PRISMA

ABSTRACT

Natural scene classification is a critical computer vision task that has shown significant advancements with the integration of Deep Learning algorithms. Traditionally, recognizing and categorizing different natural environments, such as forests, beaches, or mountains, present challenges due to the complexity and variability of natural imagery. Deep Learning has undergone a revolution with the introduction of Convolutional Neural Networks. By automatic extraction of hierarchical features from unprocessed image data, these models have surpassed traditional methods in accuracy and efficiency. These deep-learning models can capture subtle patterns and textures inherent in natural scenes, leading to more precise classifications. The current study consists of a thorough analysis of previous research on Deep Learning methods for classifying natural scenes. The publicly accessible information and data-gathering methods are used here, beginning with Web of Science, Elsevier, Springer, Scopus, IEEE Xplore, and Google Scholar, wherein hundreds of publications were selected to start the review process. Next, using a thorough screening process, seventy papers are finalized and selected for detailed literature review. The work evaluates many scene classifications based on various performance criteria. In the end, the review also paves the path for future directions for exploration in the scene classification field.



© 2025 Published by Faculty of Engineering

1. INTRODUCTION

Natural scene recognition is a challenging task in computer vision that involves automatically understanding and categorizing of real-world environments from images. This field has seen remarkable progress as the deep learning emerged and continues to be crucial for applications ranging from autonomous navigation to content-based image retrieval. We know that not every sight in nature fits

easily into the same broad category; some photographs are more representative of their category than others.

Human beings can swiftly and effectively classify short given photographs of natural sceneries, whether the category is determined by an item inside the scene or by the quality of the whole scene (Torralbo et al., 2013). Likewise, the brain imaging studies have shown that pictures of natural scenes can have differential activation in processing.

¹ Corresponding author: Priti
Email: prutigrover416@gmail.com

Numerous studies have been conducted on scene image analytics, which includes clustering, classification, and scene representation. The close relationship between scene image analytics and new technologies like robots, sensors, IoT and video cameras (Nweke et al., 2018) is evident in many applications such as satellite image analytics (Neupane et al. 2021), hyper spectral image analytics (Shadman et al., 2019), climatic image analytics (Dutta et al., 2013) and among others. As the methods used to represent the images vary depending on

their characteristics (Sitaula et al., 2019) hence researcher need to use the right feature extraction techniques for each kind of picture representation. The research in this field have advanced from conventional computer vision-based approaches to more sophisticated levels that use search engine-based methods and the most recent Deep Learning (DL) techniques (Sitaula et al., 2024). Figure 1 presents some natural scene sample images of Streets, Building, Mountain, Glaciers, Sea, and Forest.

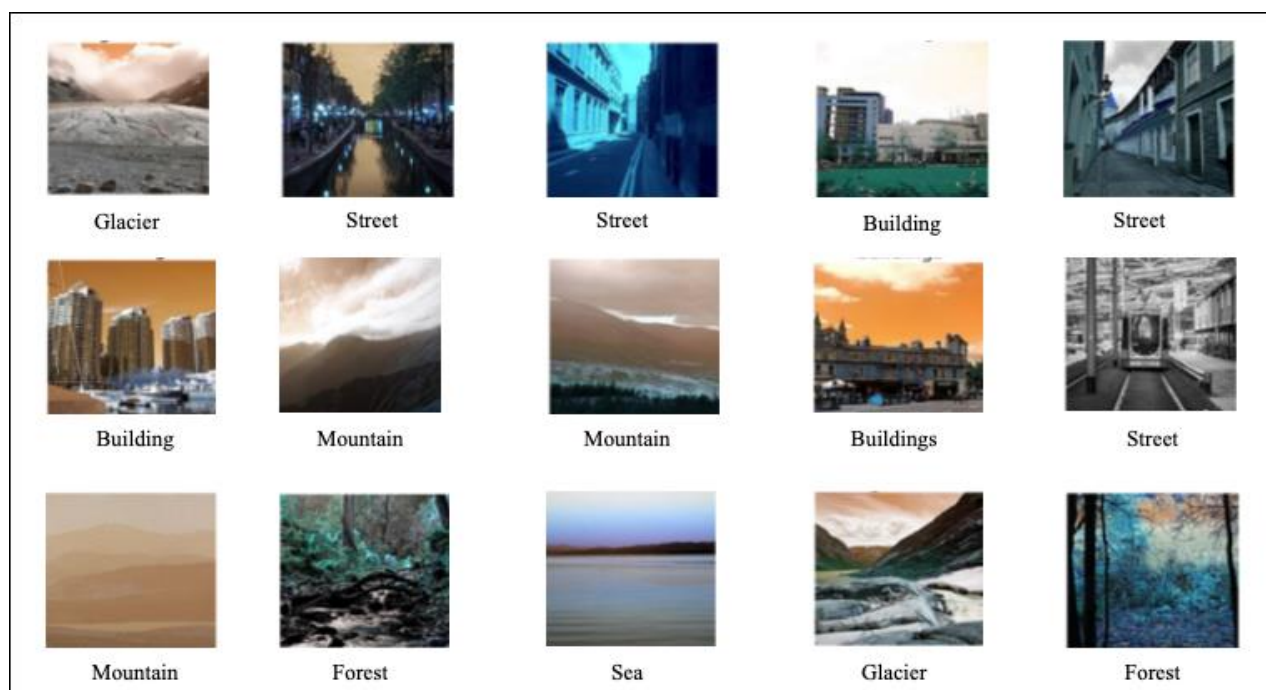


Figure 1. Example of natural scene

Photographic pictures and synthetic images are two main types of images. This hierarchy of the scene categories is described in Figure 2. Drawings, cartoons pictures, screenshots from games and other applications, logos, charts, and maps are all examples of synthetic pictures. Indoor scenery (Sitaula et al., 2019) and outdoor scenery are two further categories into which photographic images could be subdivided. Home, Recreational, Commercial, Workplace, and Public Spaces are the five main types of indoor scenes. The living room, bedroom, kitchen, bathroom, garage, stairs, children's room, foyer, and dining room are all examples of home scenes. Leisure establishments include things like buffets, restaurants, fast food joints, movie theaters, gyms, gaming rooms, hair salons, and buffets. There are many different types of workplaces, including hospitals, classrooms, art studios, laboratories, offices, schools, operating rooms, television studios, and computer rooms. Stores include things like grocery stores, jewelry

Stores, flower shops, shoe stores, malls, and video stores. Some examples of public spaces are churches, Buses, airports, waiting rooms, subways, jail cells, elevators, and train stations.

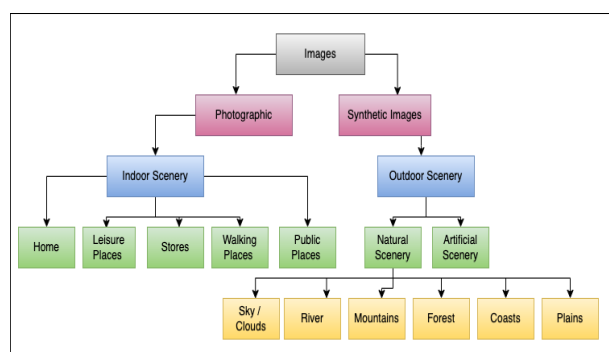


Figure 2. Hierarchy of Scene Categories (Patel et al., 2020)

Natural scenery and artificial scenery are two other categories of outdoor scenery (Raja et al., 2013). The sky/clouds, rivers, mountains, forests, and plains/deserts are the many subcategories of natural scenery. It is called artificial scenery if the scenery includes any man-made items. Examples of artificial scenes include city streets, markets, and bus stations. Hybrid sceneries are photos that include elements of both indoor and outdoor

settings. Instances of a hybrid scene include photographs of corridors.

Machine learning subfield popularly known as Deep Learning uses representation learning which means automatic extraction of meaningful features from raw data through multiple hidden layers (Lecun et al., 2015) of processing and also makes extensive use of training data for creating AI applications like classification or prediction. Computer vision (Quattoni et al, 2009) bioinformatics, automated control, machine translation, autonomous driving (Chenyi et al., 2015), Health monitoring and other real-world issues have all found optimal solutions using DL (Zhao et al., 2017). The success of DL comes from its distinct network architecture. The ability to learn small details using local features and synthesizing them at higher layers allows DL based models to understand complex features making it popular for computer vision task. DL automatically generates components by understanding complicated features from raw data using multi-level abstraction. To learn more and more complex features more quickly and effectively, DL advocates for a more in-depth model structure, emphasizes feature learning, and suggests several methods for dense scene analysis (Zhang et al., 2020). DL can adapt quickly to various situations and solve various difficulties because of their exceptional learning capacity. Both the macro-world and the micro-field have essential tasks to do while examining films or photographs containing several fascinating items, such as studies on microorganisms (Song et al., 2018) and crowded traffic. There are numerous real-world applications of DL which are solved by existing methods.

The present work is structured in eight sections. The overview has been provided for Natural Scene Recognition (NSR) and its classifications. Deep features and handcrafted features, as well as how deep learning methods contribute to the advancement are explained in section 2. Section 3 elaborated on the methodology employed in the present study wherein details of databases, keywords, selection criteria and the standards used to choose relevant study have been explained in detail. Section 4 and section 5 highlight some renowned work in scene classification using Deep Learning and Machine Learning field. Section 6 described the details of various databases available for scene classification followed by section 7 details about the findings of the paper. At the end conclusion and some direction for future work is explained in section 8.

2. SCENE CLASSIFICATION

Scene categorization differs significantly from standard object classification. Scene categorization (Nadeem et al., 2019) photos feature multiple items of various categories and spatial layouts. It plays a crucial role in real world applications where understanding of environment is essential. Scene classification provides a

broader context (e.g., highway, park, city street), making it essential for holistic understanding. The challenging work of scene categorization is further complicated by factors such as the variety in size, cluttering, occlusion, inter-class similarity (e.g:- a scene of park may look like garden or forest), intra-class similarity (a street scene be urban or suburban, crowded or empty taken in different weather conditions) and varied view angles. Many researchers are working on scene categorization. Several new methods have emerged in recent years. A few methods operate worldwide, while others focus on more specific local features and characteristics. In addition to these, mid-level representation becomes a viable option for classifying scenes as shown in Figure 3.

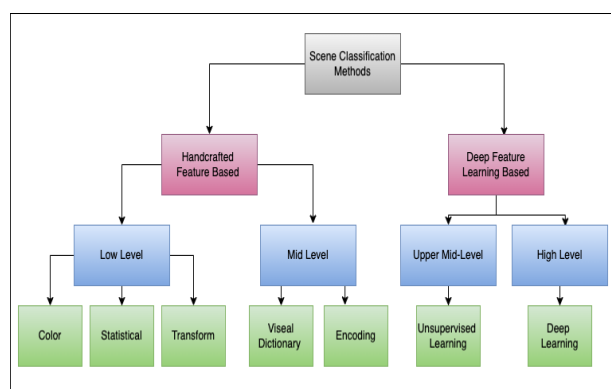


Figure 3. Types of Scene Classification Techniques (Chavda et al., 2023)

Scene classification methods are broadly classified into handcrafted feature and deep features. Handcrafted features involve manually designed features to capture key characteristics like texture, color, shape, structure. In contrast, in deep features instead of manually selecting features here more robust feature extracted by neural network from raw images.

2.1 Handcrafted Feature

Handmade or Handcrafted feature (Ahlawat et al., 2020) based techniques for scene classification involve manually selecting and designing features that are used to categorize different scenes in images. These methods rely on human expertise to identify the most relevant features that can effectively distinguish between various scene types, such as urban, rural, indoor, or outdoor settings. The procedure usually starts with feature extraction, which identifies particular aspects of the image such as color, texture, edges, and patterns. For example, methods such as the Histogram of Oriented Gradients (HOG) can be used to capture edge information, while color histograms or Gabor filters could be employed to analyze color and texture. Scale-Invariant Feature Transform (SIFT) features extract features which are robust to rotation and illumination changes. Alternatively, Generated Inputs Sets Transferability (GIST) descriptors provide a low-

dimensional representation of scene structure, which are useful in scene recognition applications.

Handcrafted feature-based methods depend heavily on the choice of features and the ability of the classification model to utilize them effectively. While these methods have been successful in various applications, they can be limited by their reliance on human intuition and can struggle with complex scenes where feature selection is more challenging. With the advent of DL, these methods are increasingly being replaced by automatic feature extraction techniques, which can learn more nuanced and complex patterns directly from the data.

2.2 Deep feature learning-based methods

Deep learning based methods utilize deep feature for learning of scene classification to automatically recognize and extract important elements from pictures that can be used to categorize different scenes. These methods leverage layers of abstraction in neural networks can extract intricate features and patterns from the visual data. Typically, a CNN is employed, where the network is trained on large datasets to recognize various scene categories by learning hierarchical features—from simple edges and textures to more complex shapes and objects. Once trained, the network can effectively classify new scenes based on the learned features, achieving high accuracy even in diverse and challenging environments. This approach reduces the need for manual feature engineering and improves performance by utilizing the deep network's ability to generalize from extensive data. Table 1 describes methods for image scene classification.

Table 1. Image scene classification methods (Shafaey et al., 2020).

Handcrafted Feature method	Deep learning Feature method
Color histogram	Stacked Auto-Encoder (SAE)
Texture descriptor (Shalaby et al., 2014)	Convolutional Neural Networks (CNNs) (Krizhevsky et al., 2012; Simonyan et al., 2014)
GIST (Generated Inputs Sets Transferability)	Deep Belief Nets (DBN)
SIFT(Scale Invariant Feature Transform) (Salem et al., 2017)	Deep Belief Machines (DBM) (Salakhutdinov et al., 2010)

3. REVIEW METHODOLOGY

The information regarding planning and selection criteria which were used to pick relevant papers for this review is presented in this section.

3.1 Search Strategy

Systematic Literature Review's (SLR) primary goal is to enhance the classification of natural scenes. The present review work is compiled using the key academic literature databases: - IEEE Explore, Science Direct, Sprinkler, Scopus, Web of Science, Study Gate, etc. The literature suggests some learning of earlier studies in the field of scene classification, related to NSR that is used by the researchers in their SLRs.

Table 2. Keyword Searching Strategy

Sr No.	Keyword
1.	What are most common challenges encountered in natural scene classification using DL?
2.	How DL techniques improve accuracy of natural scene classification compared to conventional methods?
3.	What is the role of CNNs in natural scene classification?
4.	What are some common datasets used for natural scene classification?
5.	What are the benefits of using pre-trained models like VGG16, Res-Net, or Inception for natural scene classification?

The extensive research is carried out with the objective of utilizing a variety of search keywords. The search has been performed across some databases, including Scopus, Sprinkler Link, IEEE Xplore, and Science Direct irrespective of its publication date. Search parameters include keywords from the topic, abstract, and title. Only particular categories of documents, including bibliographies, proceedings papers, conference papers, reviews, and articles, were taken into consideration. Table 2 lists the keyword-search approach that was used for this study and Table 3 represents how many papers are retrieved for each keyword.

Table 3. Total number of papers retrieved for each keyword.

List of Searched Keywords	Paper Extracted
DL	50
Natural Scene	65
ML	40
Remote Sensing Image	55
CNN	49
Total	259

The keywords which are used for our paper are shown in Figure 4 using graphical format.

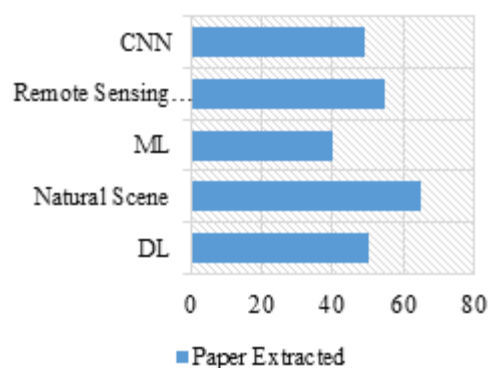


Figure 4. List of various keywords

A total of 259 papers were retrieved via the use of various keywords, and then inclusion and exclusion procedures were carried out on those articles.

3.2 Selection Criteria

Prior studies are done to determine how frequently the most well-known keyword search for each query without filtering. Data collection studies and validation/reliability studies were used to categorize the findings. After that, the papers titles, summaries, and

methodology sections were omitted from the list. This was done to determine what other terms were utilized in the trials and to exclude irrelevant publications from the study. Then we created a list of terms and issued separate inquiries for each one. In the end, eleven additional individuals were included. We summarized the results and listed the keywords that were used to search for each item.

3.3 Scrutinizing of paper for study

Our exploration highlights selection procedure for the primary studies in four stages: detection, admissibility, inclusion, and multiple screening. Through a systematic search of several sources, including full-text articles, IEEE Explore, Science Direct, Sprinkler, Research Gate etc., 18,300 searches were found in the first phase. Table 4 shows criteria for inclusion and exclusion of articles. In the second phase, a preliminary evaluation was performed by looking over the article's title, keywords, and abstract. These two documents with the notation combined and unclear were sent for further investigation. Figure 5 shows PRISMA-based flowchart for systematic review of publications found in databases.

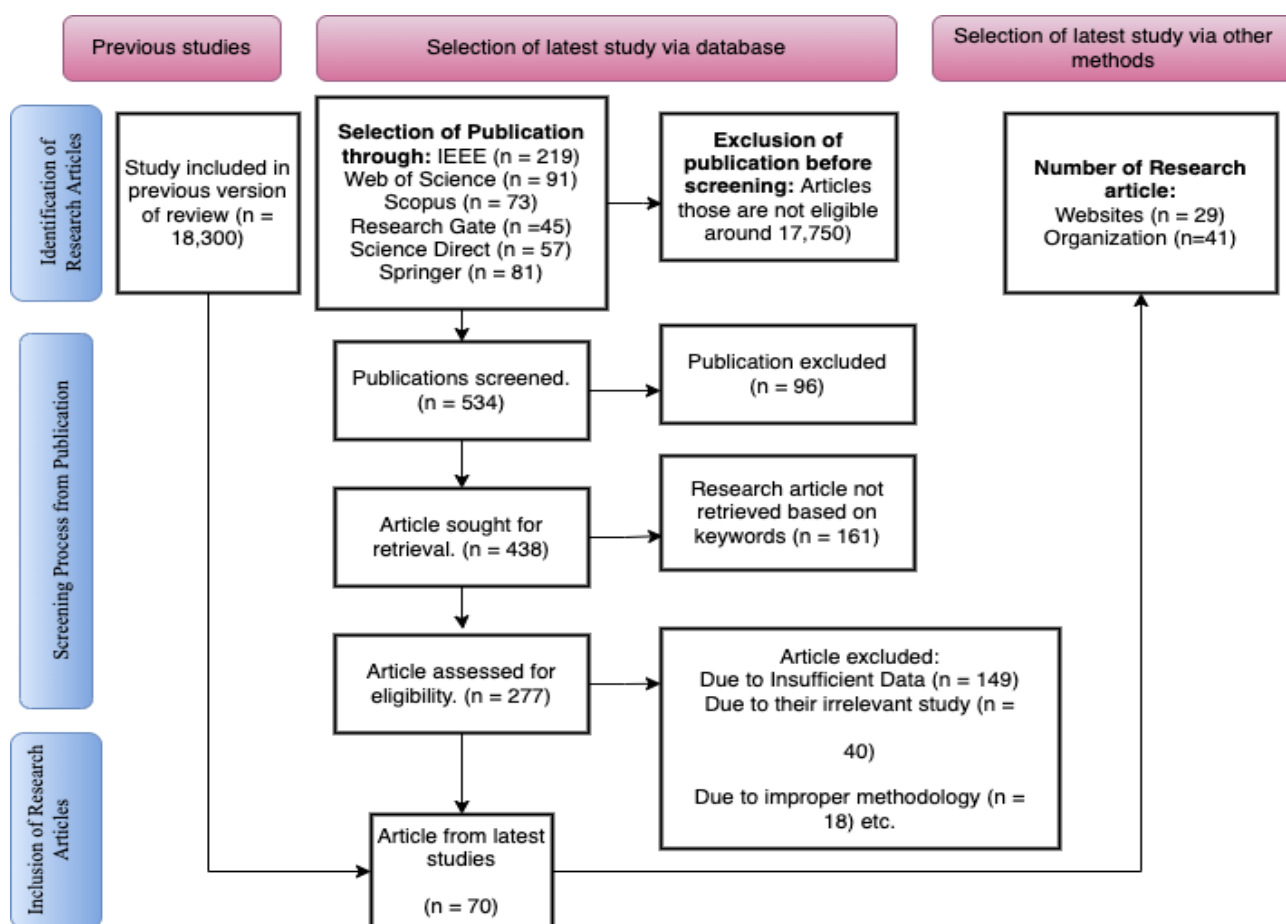


Figure 5. PRISMA based Flowchart for Systematic review of Publications

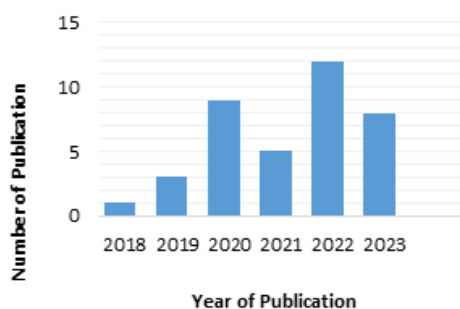
Table 4. Criteria for determining Inclusion/Exclusion of Articles.

Criterion	Inclusion	Exclusion
Keywords	Publication conferring the relationship	Publication excluded in which variables have no relation
Type of Literature	Journals, Review Articles	Book series, book, chapter in book,
Language	English	Other than English
Timeframe	Concerning 2012-2024	<2012
Category	Open Access	Paid Access

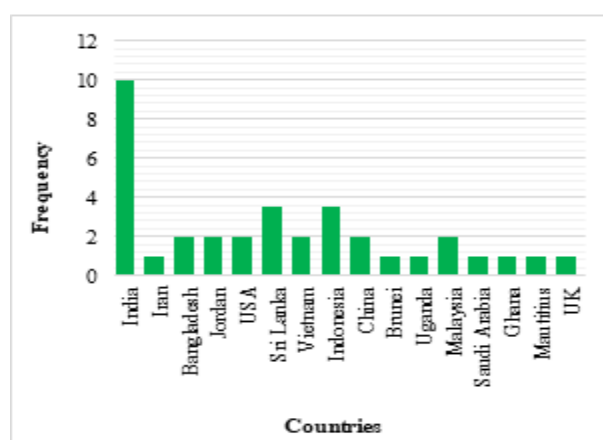
Then we depict the comprehensive systematic process of extracting data in Figure 6. The process involves extracting research articles, conducting source analysis, excluding irrelevant material based on predefined criteria, and finally analyzing the substance of the source research papers. A specific time range of 12 years (2012-2024) has been designated for conducting a comprehensive systematic literature review to obtain specific papers on the topic area. A total of 1273 studies are screened after automation tools remove ineligible studies. After screening, articles are sought for retrieval. Out of the 534 publication, screened articles for 438 are retrieved. Articles are then assessed for eligibility. A total of 278 articles are assessed for eligibility. Out of the assessed articles, 149 are excluded due to insufficient data, 40 are excluded due to irrelevant study and 18 are excluded due to improper methodology. Finally, a total of 112 articles of new studies are included. Additionally, the remaining papers underwent domain-based screening to determine if the retrieved research papers align with the variables of the studies. The total number of publications that met this criterion was (n=70). The retrieved papers underwent a screening process to determine the language in which they were written, as well as to identify any duplicate papers. The study focused exclusively on works written in English.

3.4 Demography of publication

The analysis shows study, which aimed to examine the changing interest in Natural scene classification and identification by aggregating data from research papers based on their publication years in Figure 6. According to the data, there will be a large increase in publications between 2018 and 2023. As seen by the steady increase in the quantity of papers published in this field during the past few years, these results indicate a rising and continuous interest in this field of study.

**Figure 6.** Number of articles published annually

An examination of the connections of the initial authors was carried out so that they could get a full grasp of active research groups and the geographical dispersion of such organizations. Figure 7 is a graphical representation of International distribution of research article.

**Figure 7.** International research distribution

4. LITERATURE REVIEW ON SCENE CLASSIFICATION USING DEEP LEARNING

Here, a comprehensive examination of DL methods for scene classification. A CNN is the DL (Lecun et al., 2015) architecture that is most often used. The typical pipeline of most CNN-based approaches for classifying scenes is displayed in Figure 8 using CNN as the feature extractor. The first step in almost any scene picture input problem is to employ CNN extractors to acquire local information. The next step is to process images by applying filters to find out major features using encoding then concatenating these features through multiple layers applying pooling. Then combining these characteristics into a representation at the picture level. Then taking representation as input we obtain a category which is the last step in the classification process, on the basis of features we label category of the image.

Dorrani et al. (2024) analyze traffic using new DL technologies with CNNs. A proposal has been made to use ResNext architecture, which is very effective and generate interest from many other areas, to analyze the scene and separate it into three categories: people, bikes

and cars. The suggested strategy outperforms its competitors according to important assessment parameters including sensitivity, specificity, and accuracy. The sensitivity level was about 89%, while the accuracy criteria were up to 90.25%.

Naseer et al. (2024) introduced a fresh method for precise object identification. Object recognition in a variety of settings is the ultimate goal of CNN technology. The strategy's efficacy is shown by

performance measures for MSRC, the F1 Score is 0.948, the Precision is 0.968, and the Recall is 0.932. For Caltech 101, the F1 Score was 0.921, the Precision was 0.951, and the Recall was 0.891; for Pascal Voc 2012, the F1 Score was 0.847, the Precision was 0.879, and the Recall was 0.827. Using multiple benchmark datasets, such as Pascal Voc 2012, Caltech 101, and MSRC-v2, the proposed technique is tested and shown to be effective with 92.25%, 91.91%, and 93.50% recognition accuracy, respectively.

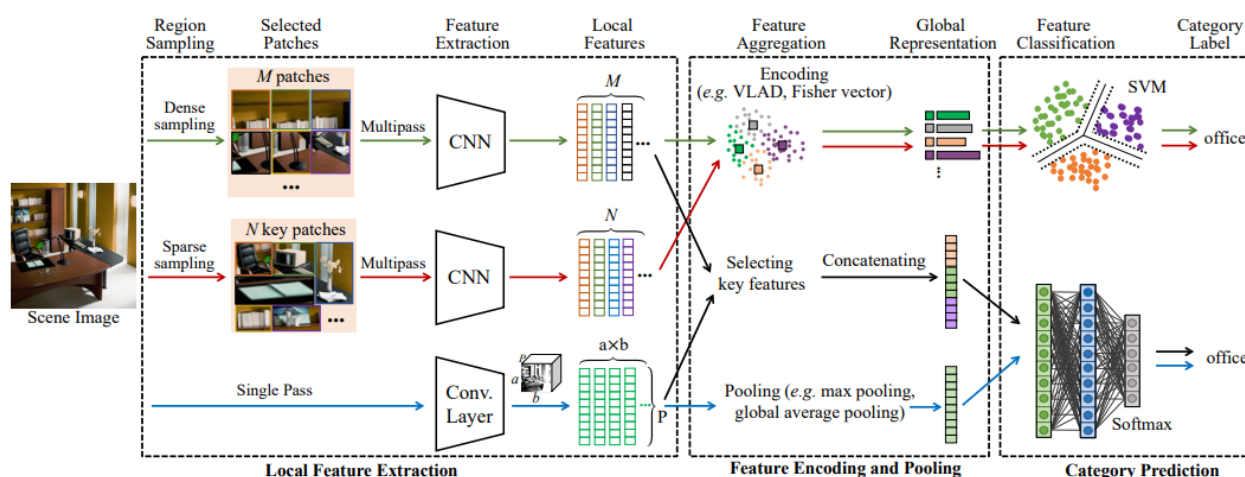


Figure 8: The generic DL pipeline for scene categorization (Zeng et al., 2021)

Sitaula et al. (2024) offered a new DL-based method that makes use of an improved VHR attention module (EAM) to classify VHR RS images more accurately by Capturing more salient multi-scale information. The suggested method achieves the lowest standard deviation of 0.001 in classification performance, which is both competitive and stable/consistent, according to experimental findings on two popular VHR RS datasets. In addition, the AID dataset has the best overall accuracy at 95.39%, while the NWPU dataset has the best overall accuracy at 93.04%.

Oktavian et al. (2024) combined four different CNN layers—InceptionV3, Xception, MobileNetV2, and ResNet50—to produce a new ensemble learner that aims to surpass the best results of the individual layers. The Intel Image Classification dataset, consisting of six fully coloured nature sceneries for benchmarking. The results shows model has an average accuracy of 83%, while the ensemble mode achieved 88% accuracy.

Sivasubramanian et al. (2024) used vision Transformers (VT), Swin Transformers (ST), and ConvNeXt as an ensemble method of feature extraction that is inspired by federated learning. A dataset comprising 36,785 photos from 92 classes, after merging all subclasses from five reference datasets, has been used for remote sensing natural scenes. The experimental results with the suggested

method achieved a high F1-score of 0.9587 and a test accuracy of 97.24%.

Pereira et al. (2024) the study proposes a novel approach for indoor scene classification that utilizes semantic segmentation based features and global features. A two-branch network namely GS2F2App comprised of segmentation-based features from suggested SSFs and integrated with CNN-based global Features that has been evaluated on SUN RGB-D and NYU Depth V2 indoor scene benchmark datasets. The proposed model achieved state-of-art performance and highlighted the hybrid feature strategy.

Akhtar et al. (2023) used transfer learning across domains to improve land-cover categorization in remotely sensed photos of landscapes. The research indicates the classification accuracy using ResNet-50 model is higher among other models. The cross-domain transfer-learning system outperformed previous models on the NaSC-TG2 benchmarks dataset and achieved an accuracy of 99.5% with full model fine-tuning and 99.1% when fine-tuning only the last three layers.

Chen et al. (2023) suggested PyHE-Net, a general post-quantum platform that seamlessly integrates textual DL tools with encryption. Also, researcher improve the pooling, convolution, and activation

functions and accomplish the first-ever cipher-text operation under floating point values. The findings demonstrate that, in comparison to other cryptographic models, PyHENet is both sufficiently resilient and closest to the plaintext inference platform. The enhanced PyHENet algorithm achieved 95.05% accuracy in the high-resolution NaSC-TG2 database.

Biswas et al. (2023) developed an innovative DL model to classify stunning photos of natural scenes by fusing statistical and semantic data. This method yields average classification rates of 92.74% on the dataset, 88.67% on the AVA dataset, and 83.22% on the CUHKPQ dataset, experiments conducted on the dataset and the standard datasets. The suggested approach is successful for the categorization of beautiful social media photographs, according to a comparison study of the suggested model with the current approaches.

Tang et al. (2022) introduced a novel Context Aggregation Module (CAM) consisting of two main components: gated fusion and context encoding. The work balances different-scale contexts during feature fusion using an expanded gate attention mechanism. The method is tested on challenging datasets mainly Cityscapes, CamVid, and BDD100K, to evaluate autonomous driving scenarios and reported significant improvement in urban scene segmentation area.

Liu et al. (2022) replaced four layers of VGG-19 model with two self-attention blocks comprised of two multi-head self-attention layers also known as MHSA for remote sensing imagery. The proposed approach focuses on spatial regions by utilizes both regional and global data. This self-attention and CNN integrated model is evaluated on four public data sets, namely Euro-Sat, NaSC-TG2, AID and WHU-RS19, and reported the usefulness of the suggested technique.

Zhou et al. (2021) detailed "NaSC-TG2," dataset which is a new standard for natural scene categorization using remote sensing data collected from the Tiangong2 satellite. There are 20,000 photos in dataset, distributed evenly across 10 different scene classifications. Tiangong-2, a state-of-the-art wideband imaging spectrometer, captured all of the photos. Along with the RGB photos, they also give the equivalent multispectral scene images. As this study shows, this dataset was great for testing and developing classification systems.

Iyer and Rishidas (2020) presented a CNN algorithm with an Alex-Net model to classify images of natural spaces. Researchers define several types of nature settings according to their characteristics and are mostly focused on 12 distinct categories here. A

dataset including 100 images of natural settings is used to train the model. The model Alex-Net is used as the neural network. Every picture categorization includes a histogram analysis. Additionally, the picture accuracy and the model loss are determined. The performance of the model is assessed using a confusion matrix that displays the actual value for every class.

Shen et al. (2020) presented RA-Net, an innovative attention network for representing the connection between areas of objects. Researcher train the object areas' construction, the representative regions' contents selection, the information pathways' setup, and the pixels' context exchange to increase segmentation accuracy. Extensive testing on difficult segmentation benchmarks shows that RA-Net works efficiently.

Laulkar et al. (2020) provided technique for classifying photographs of outdoor nature scenes based on predetermined semantic criteria. By using SPDBSCAN in conjunction with human input on the source picture, segmented images could be produced. Using these segmented photos, CNN model is trained to recognize things like sand, water, sky and green land. Semantic rules are developed based on the item class and its spatial location in an image to classify scene photographs into four different classes: beach, seawater, desert, & green ground. The SUN-397 dataset provided the photos utilized in this study. A scene categorization F-ratio of 84% was achieved in the study.

Jayanth (2020) used a CNN algorithm to categorize sceneries from nature into six distinct groups. As a penultimate step, this model estimates the likelihood of various class labels, which is then used to make predictions about the final class. There are two main uses for this dataset: training and testing. It delivers an accuracy of 84.93%. Due to the uncertainty introduced by images including elements from both situations, the model has a hard time while classifying.

Table 5 depicts comparison of Literature via Deep Learning and Figure 9 represents the accuracy achieved on various literature articles via Deep Learning.



Figure 9. Comparison of Accuracy mIoU

Table 5. Comparison of Literature Review based on DL.

Authors (Year) (Reference)	Scene image Classification	Dataset Used	Methodology	Outcomes
Dorrani et al., (2024)	Traffic scenes	—	ResNeXt	Categorization accuracy for vehicles stands at 92.4%, whilst pedestrians only manage 78.2%. Overall, the combined technique has a much higher accuracy rate of 90.2%.
Naseer et al., (2024)	Outdoor Natural Scenes	MSRC-v2, Caltech 101, and Pascal Voc 2012	CNN	The usefulness of the suggested approach was demonstrated by the 92.25%, 91.91%, and 93.50% recognition accuracies that were obtained using the MSRC-v2, Caltech101, and Pascal Voc 2012 benchmark datasets, respectively.
Sitaula et al., (2024)	Remote Sensing Scene	AID and NWPU	EAM	On the AID dataset, the suggested technique boasts an accuracy of 95.39%, but on the NWPU dataset, it drops to 93.04%.
Oktavian et al., (2024)	Multi-class image	Intel Image Classification	InceptionV3, Xception, MobileNetV2, and ResNet50	The ensemble mode achieved an impressive 88% accuracy on the test set, even though each model only achieved 83% accuracy on average.
Sivasubramanian et al., (2024)	Remote sensing images	VT+ST+Conv NeXt	RS_NS92	The experimental results with the suggested method not only achieve a high F1-score of 0.9587 and a test accuracy of 97.24%.
Pereira et al., (2024)	Indoor scene classification	NYU Depth V2 and SUN RGB-D	GS ² F ² App	Two top-notch interior scene datasets, NYU Depth V2 (accuracy of 77.8%) and SUN RGB-D (62.3%), were used to evaluate the GS2F2App.
Akhtar et al., (2023)	Remote sensing images	NaSC-TG2	ResNet-50	The proposed model has 99.50% accuracy, recall, and precision.
Chen et al., (2023)	Remote sensing images	NaSC-TG2	PyHENet	The proposed model attains an accuracy of 95.05% on the training set but falls to 90.43% on the testing set.
Biswas et al., (2023)	Aesthetic natural scene images	Own, AVA, and CUHKPQ dataset	DL	The suggested method yields average classification rates of 92.74%, 88.67%, and 83.22% on the proposed dataset, according to experiments conducted on both the own and standard datasets.
Tang et al., (2022)	Urban scene	Cityscapes, Cam-Vid, and BDD100K	CA-Net	On Cityscapes, the suggested model achieves a competitive mIoU of 82.7% while on Cam-Vid, it achieves an ideal mIoU of 80.5%.
Liu et al., (2022)	Remote sensing images	NaSC-TG2, WHU-RS19, AID and Euro-SAT	VGG-SA	The proposed model has a 98.57% accuracy, recall, and precision.
Zhou et al., (2021)	Natural Scenes	NaSC-TG2	Google-Net	The accuracy level of the suggested model is 87.76%.
Iyer and Rishidas (2020)	Natural Scenes	Natural scene dataset	Alex-Net	An accuracy rate of 89.6% was achieved with the Alex-Net model.
Shen et al., (2020)	—	Cityscapes, Pascal Context, and COCO Stuff	RA-Net	On the Cityscapes dataset, the proposed model attains an accuracy of 81.9; on the Pascal Context dataset, it reaches 54.9; and on the COCO Stuff dataset, it reaches 40.7.
Laulkar et al., (2020)	Outdoor Natural Scene Images	SUN-397 dataset	CNN	The study successfully classified scenes with an 84.09% accuracy and an F-ratio of 84%.
Jayanth (2020)	Natural Scenes	Natural scene dataset	CNN	The accuracy rate provided by the proposed model is 84.93%.

5. LITERATURE REVIEW ON SCENE CLASSIFICATION USING MACHINE LEARNING (ML)

This section will offer a thorough examination of the various machine learning methods available for scene classification. Here Figure 10 illustrates various types of machine learning that could be applied to scene classification.

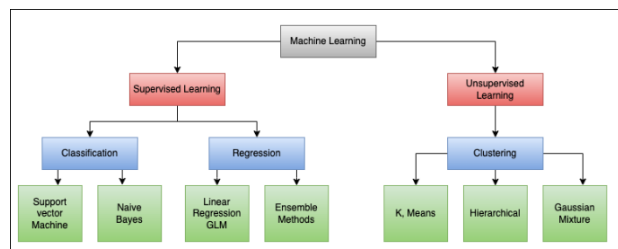


Figure 10. Different kinds of machine learning for scene classification (Patel et al., 2018)

Yang et al., (2024) explored large datasets of unlabeled photos and suggested a semi-supervised method for scene categorization in remote sensing images using prototype-based consistency. First, they suggest a feature augmentation module for extracting discriminative features; this would help us achieve the goal. They perform several analyses using the NWPU-RESISC45 dataset and AID –an aerial image Dataset. The latest method improved accuracy on NWPU-RESISC45 dataset from 92.03% to 93.08% and on the AID dataset from 94.25% to 95.24%.

Ganashree et al., (2024) presented a CNN based model named as IABC-CNN model as an improved method for enhancing artificial bee colonies for improved land scene classification. The photos are from databases from the UCM, AID, and NWPU-RESISC45. Using VGG-16 which is visual geometry group-16, IABC efficiently chooses the most relevant characteristics from the retrieved features. The classification procedure utilizes the characteristics picked from the IABC and is supported by a multiclass-support vector machine (MSVM). Classification accuracy improved to 96.40% with an error rate of 3.6% according to results collected from the planned IABC-CNN.

Shao et al., (2024) presented an event extraction block by obtaining the exact semantic segmentation area from the segmentation network, that can replicate the effect of picture segmentation on audio characteristics. The Acoustic Scenes and Events (DCASE) 2020 ASC Task 1A dataset has been evaluated and proposed model reported an average accuracy of 71.26%. They make use of the official challenge dataset, and the model is more straightforward and more effective than other cutting-edge CNN algorithms.

Amirat et al., (2024) presented a novel technique for scene categorization that utilizes the Belief Function Theory (BFT) for managing uncertainty information more effectively than conventional probability-based approaches. Researcher test the algorithms on 3 renowned datasets —SUN397, MIT Indoor, and Label-Me—and evaluate them against the most recent methods. Using a simple and resilient structure, the technique beats previous approaches in some instances while achieving competitive outcomes.

Kumar et al., (2023) implemented a novel model to enhance the capability of feature extraction for enhanced scene categorization. Aerial photos acquired from UCM, AID, and RESISC45 are processed using a multiscale Retinex method to restore colors and boost contrast. An overall classification accuracy of 99.14% on the UC Merced dataset is reported and accuracy of 98.09% on the RESISC45 dataset and 99.25% on the aerial photo dataset is also observed.

Ran et al., (2023) proposed a method for classifying scenes using a multi-scale CNN with a whale optimization algorithm (WOA) and long short-term memory (LSTM). To display the performance of the suggested method, which obtained a 98.87% classification accuracy using real-world scene data consisting of eight rooms, ablation experiments were conducted. Also, the model is tested on FR079 public dataset and observed classification accuracy of 94.35% which is better than the most advanced algorithms.

Soudy et al., (2022) here innovative few-shot learning model-Generic-Conv- presented to classify scenes, which has been tested on benchmarked datasets like Mini-Sun, Mini-Places, and MIT-Indoor 67. The suggested model Generic-Conv achieves accuracies of 52.16 ± 0.015 , 35.86 ± 0.014 , and 37.26 ± 0.014 for five-shots on Mini-Sun, Mini-Places, and Indoor 67 datasets from MIT, respectively.

Hao et al., (2022) presented an ML approach that measures the prevalence of seven different kinds of sound sceneries using CNNs and Target Sound Area Ratios (TSARs). Nine locations in three Guangzhou, China, urban woodlands were monitored for acoustic data. The work observed an F1 score of 0.97, a recall of 0.97 and a precision of 0.96 after training the CNN using the site-specific audio data.

Hasan et al., (2022) demonstrated a Genetic Algorithm (GA) that efficiently searches for CNN architectures that are optimal for the ASC challenge. It was found that the defined approach outperformed the baseline systems on the DCASE2018-Task1A, DCASE2019-Task1A, and DCASE2020-Task1A development datasets by about 17.8%, 16%, and 17.2%, respectively, based on relative accuracy. The

results on all three benchmark datasets show the effectiveness of Genetic algorithm in the proposed approach.

Shi et al., (2022) presented a model for scene categorization using deep CNNs that account for the visual focus area's depth function. Experiments are conducted on four typical scene datasets: LabelMe (OT), UIUC-Sports (SE), Scene-15 (LS), and MIT67 (IS). The findings of the suggested approach study demonstrate that it is both flexible and effective in enhancing classification performance.

Figure 11 represents the comparison based on accuracy for various Machine Learning methods.

A comparison of literature reviews based on scene categorization using ML approaches is included in table 6 as shown below.

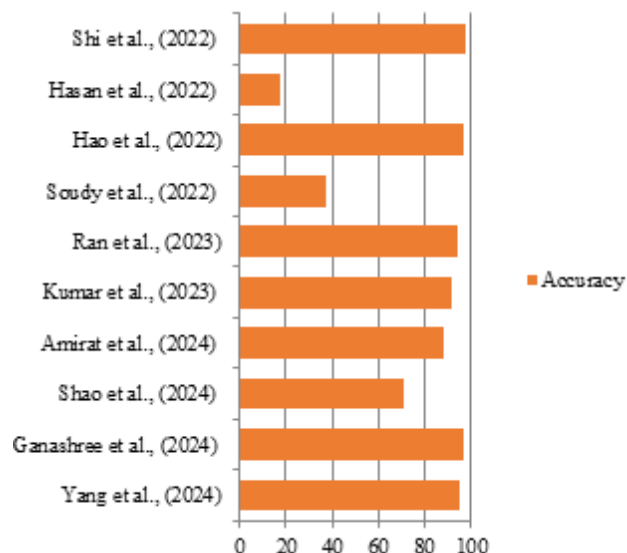


Figure 11. Comparison of accuracy based on the ML method

Table 6. Comparison of Literature Review based on ML.

Author (Reference)	Year	Dataset Used	Methodology	Outcomes
Yang et al., (2024)	2024	NWPU-RESISC45 and AID	DNN	The technique's accuracy varies from 92.03% to 93.08% on NWPU-RESISC45 and from 94.25% to 95.24% on AID.
Ganashree et al., (2024)	2024	AID, NWPU-RESISC45 and UCM	IABC-CNN	Classification accuracy improved to 96.40% with an error rate of 3.6% according to results collected from the planned IABC-CNN.
Shao et al., (2024)	2024	DCASE 2020	Mini-Seg-Net	On the various devices included in the development dataset, the suggested solution had a 71.26% success rate.
Amirat et al., (2024)	2024	SUN397, MIT Indoor, and Label-Me	BFT	On MIT, the suggested strategy improves classification accuracy to 53.68%; on SUN397, it rises to 58.98%; and on Label-Me, it reaches 87.8%.
Kumar et al., (2023)	2023	UCM, AID, and RESISC45	UFS-LSTM	The suggested model attained an overall classification accuracy of 99.14% on UCM, 98.09% on RESISC45, and 99.25% on AID.
Ran et al., (2023)	2023	FR079 public dataset	WOA	This research proposes a method for robot room categorization that achieves an accuracy of 94.35%.
Soudy et al., (2022)	2022	Mini-Sun, Mini-Places, and MIT-Indoor 67 datasets	Generic-Conv	In five-shot five-way classification, the proposed model outperformed all previously published models with an accuracy of 37.26 ± 0.014 .
Hao et al., (2022)	2022	Site-related sound data	CNN	A recall of 97%, accuracy of 96%, and F1 score of 97% were all accomplished by the suggested model.
Hasan et al., (2022)	2022	DCASE2018-Task1A, DCASE2019-Task1A, and DCASE2020-Task1A	GA	Accuracy rates of around 17.8%, 16%, and 17.2% were reached on the three datasets that were suggested.
Shi et al., (2022)	2022	LabelMe, UIUC-Sports, Scene-15, and MIT6	Deep CNN	The suggested strategy improves upon previous results in terms of classification accuracy, reaching 97.70% on the OT dataset, 93.75% on the SE dataset, 96.01% on the LS dataset, and 72.37 on the IS dataset.

6. DATASETS USED FOR NATURAL SCENE CLASSIFICATION

Here we have described the most popular data sets used by already performed research that fell within the purview of our analysis. General, interior, outdoor, and street scene datasets are classified in Table 7. In addition, the dataset's class counts, picture counts, training/validation/testing splits, and image resolutions are all detailed in this table. All three phases—training, validation, and testing—use different sets of pictures. This facilitates efficient model training and evaluation for researchers and developers. A graph comparing datasets according to their class count is shown in Figure 12.

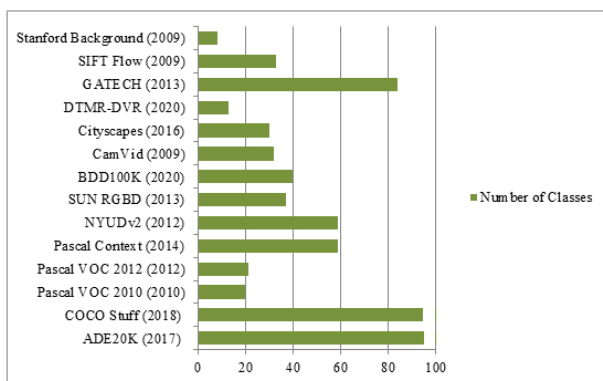


Figure 12. Comparison graph of datasets

- **ADE20K**

ADE20K is an extensive and varied dataset that includes more than 20,000 photos of both indoor and outdoor scenes annotated at the pixel level. Things like walls, buildings, skies, people, roads, beds, and countless more are included in the categories. This dataset was used to build a benchmark for scene parsing with 150 object and stuff classes. They have meticulously labeled all of the photographs with items, objects, and object pieces. The object's opacity, cropping, and other attributes have been detailed further. Part annotations in the validation set are more thorough than in the training set.

- **COCO Stuff**

The COCO-stuff dataset is an expansion of the original COCO dataset that includes extensive pixel-wise thing annotations (Lin et al., 2014). Things and their stuff could be investigated in great detail with the help of COCO-Stuff (Caesar et al., 2018). This dataset contains 172 classes, comprising 80 items, 91 items, and 1 unlabeled class. Just as in the original COCO, there are 80 item classes. A professional annotator has selected the 91 types of objects. The unlabeled class is used when a label does not match one of the 171 predefined classes or when the annotator is unable to identify the label of a pixel.

Classes for things include objects like bears, bicycles, stop signs, knives, people, parking meters, clocks, traffic lights, and so on. Numerous non-object categories, including materials, places, and more, make up the stuff classes. For instance, filth, mist, hair, dots, plastic, grid, screen, etc. By making this dataset public, the authors hoped to encourage more research into the contextual links between things and stuff.

- **Pascal VOC (Visual Object Classes)**

The Pascal VOC challenge ran from 2005 to 2012, with yearly dataset updates thereafter. When it comes to the semantic segmentation problem, it is the dataset that is most often used in the literature. The collection includes a variety of animal classifications, including horses, lambs, cows, dogs, cats, and birds. Televisions, sofas, dining tables, potted plants, bottles, and chairs are examples of indoor items. Automobiles, motorbikes, trains, buses, bicycles, boats, and airplanes are examples of vehicle classifications. Furthermore, by assigning the 'Void' label to problematic items, they were eliminated from the training and test sets. Thus, there are now a total of 21 classes in Pascal VOC 2012 (Everingham et al., 2012).

- **Pascal Context**

The previous Pascal VOC 2010 dataset was expanded into this one, called Pascal Context. Pascal Context (Mottaghi et al. 2014) differs from the PASCAL VOC 2010 dataset in two significant ways: the quantity of classes and the level of annotation detail. Pascal VOC supports annotating about 20 different kinds of objects. As a counterpoint, Pascal Context's 540 classes allow for significantly more detailed picture annotations, including 80 thing classes, 459 stuff classes, and one unlabeled class. The photographs have been annotated pixel by pixel. Out of all these classes, 59 were deemed the most significant. Subjects include things like birds, bottles, cows, people, clouds, floors, snow, sea, wood, and windows, among others.

- **Berkeley Deep Drive (BDD100K)**

One dataset that could shed light on the challenges of street-scene interpretation is the BDD100K dataset, which consists of large-scale, diversified driving videos annotated extensively (Yu et al., 2020). 100k video sequences shot at a smooth 30 frames per second in high definition (720p) make up the dataset. These movies are a compilation of footage shot in several locations, including the Big Apple and the San Francisco Bay Area. The collection includes scenes (e.g., residential, city street, highway, and tunnel), weather conditions (e.g., clear, rainy, snowy, and foggy), and time of day (e.g., day and night).

- **SIFT Flow**

There is a subset of the Label-Me dataset that is known as the SIFT Flow dataset (Liu et al., 2009). The collection includes 2688 completely annotated (Russell et al., 2008) photos in all. Outdoor settings with components like bridges, mountains, roads, traffic

lights, towers, and water are prevalent in most of these photos. They have determined the 33 most prevalent item types by tallying the number of labeled pixels. The unlabeled category serves as the 34th level of classification and comprises pixels that have been classed as another object type or are not labelled.

Table 7. Summary of static scene parsing datasets

Content	Dataset, Year	Number of Classes	Number of Images	Samples			Image resolution
				Train	Validation	Test	
Generic	ADE20K (2017)	150	25k	20,210	2000	3352	2400*1800
	COCO Stuff (2018)	171	164k	118k	5k	45k	Variable
	Pascal VOC (Everingham et al., 2010)	20	1928	771	289	868	500*400
	Pascal VOC 2012 (2012)	21	4369	1464	1449	1456	Variable
	Pascal Context (2014)	59	10,103	4998	—	5105	Variable
Indoor	NYUDv2 (Silberman et al., 2012)	59	10,103	4998	—	5105	Variable
	SUN RGBD (Xiao et al., 2013)	37	10,335	5285	—	5050	Variable
	BDD100K (2020)	40	100k video frames	70k	10k	20k	1280*720
	CamVid (Brostow et al., 2009)	32	701 video frames	367	100	233	960*720
	Cityscapes (Cordts et al., 2016)	30	5000	2975	500	1525	2048*1024
	DTMR-DVR (Pfeiffer et al., 2013)	13	600 video frames	400	100	100	1280*960
Outdoor	GATECH (Hussain et al., 2013)	84	20k video frames	13k	7k	7k	Variable
	SIFT Flow (Liu et al., 2009)	33	2688	2488	—	200	256*256
	Stanford Background (Gould et al., 2009)	8	725	572	—	143	320*240

7. ANALYSIS

Based on the analysis we have found major findings of paper and some key suggestions are:

- **Advancements in Deep Learning:** Our exploration highlights how deep learning, particularly convolutional neural networks (CNNs) and other methods, has revolutionized the classification of natural scenes by automatically extracting feature and achieving higher accuracy compared to traditional methods.
- **Datasets and Benchmarks:** We have Defined about how important it is to train and test deep learning models in natural scene categorization tasks utilizing sizable, varied, and openly accessible datasets like ImageNet or CIFAR-10.
- **Model Architecture Recommendations:** Our study also Emphasized on selection of suitable architectures like CNN, ResNet, or other variants that are well-suited for handling the complexity and variability of natural scenes. Also we explained how DL models tend to perform better but also

require more computational resources for their better working.

- **Challenges and Future Directions:** We have Acknowledge the limitations, such as over fitting on small datasets, and provide methods to get around them, such as data augmentation or transfer learning. On the basis of our analysis we have suggested that in future research we can explore combining deep learning with other techniques, like reinforcement learning for improving classification performance.

8. CONCLUSION

In conclusion, we have presented various DL methods for the classification of natural scenes and comparison of various algorithms and models like ML models and DL (e.g., SVM, Auto encoder, GAN, CNN, RF algorithms, etc.). Our review paper scope includes a thorough investigation of several deep learning models and algorithms used for tasks involving the classification of natural scenes. We discussed significant developments in recurrent neural networks

(RNNs), convolutional neural networks (CNNs), and other techniques. This review highlights the datasets commonly used, performance metrics, and challenges in the field. Furthermore, it discusses the impact of data augmentation, transfer learning, and computational efficiency. The literature study seeks to offer a thorough grasp of the developments and patterns in deep learning for the classification of natural scenes. The research study is done by using a variety of selection categories, keywords, and tight collection criteria.

Scene Classification models and monitoring tools based on ML and DL have shown great results. From 2012 to 2024, a total of 70 publications were included in the research.

We used six academic database and used strict inclusion criteria to choose these papers. Significant and relevant results from the research that were evaluated were given in the analysis. On this subject, publicly accessible datasets are few. Furthermore, most DL models are built using data acquired in a controlled

laboratory environment, which might impact how well they function when used in real time. The limitations of the research are:

- **Imbalanced datasets:** In natural scenes, certain classes (like urban scenes or mountains) may be overrepresented, while others (like rare biomes) might have limited data.
- **Over fitting to training data:** Overly complicated models could perform well on training data but poorly generalize to unseen images.
- **Energy consumption:** Related to the computational demands, there are many concerns about the environmental impact and sustainability of training large deep learning models.
- **Scalability to different scenes:** Models trained on specific types of natural scenes might struggle when they are applied to different geographic locations or environmental conditions.

References:

- Ahlawat, S., Choudhary, A., Nayyar, A., Singh, S., & Yoon, B. (2020). Improved Handwritten Digit Recognition Using Convolutional Neural Networks (CNN). *Sensors*, 20(12), 3344. <https://doi.org/10.3390/s20123344>
- Akhtar, M., Murtza, I., Adnan, M., & Saadia, A. (2023). Cross-domain transfer learning for natural scene classification of remote-sensing imagery. *Applied Sciences*, 13(13), 7882. <https://doi.org/10.3390/app13137882>
- Amirat, A., Benrais, L., & Baha, N. (2024). Towards exploiting believe function theory for object based scene classification problem. *Multimedia Tools and Applications*, 83(13), 39235-39253. <https://doi.org/10.1007/s11042-023-17120-z>
- Biswas, K., Shivakumara, P., Pal, U., Lu, T., Blumenstein, M., & Lladós, J. (2023). Classification of aesthetic natural scene images using statistical and semantic features. *Multimedia Tools and Applications*, 82(9), 13507-13532. doi: 10.1007/s11042-022-13924-7
- Brostow, G. J., Fauqueur, J., & Cipolla, R. (2009). Semantic object classes in video: A high-definition ground truth database. *Pattern recognition letters*, 30(2), 88-97. doi:10.1016/j.patrec.2008.04.005
- Caesar, H., Uijlings, J., & Ferrari, V. (2018). Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1209-1218). doi: 10.1109/CVPR.2018.00132
- Chavda, S., Goyani, M. Scene Level Image Classification: A Literature Review. *Neural Process Lett* 55,2471–2520(2023). <https://doi.org/10.1007/s11063-022-11072-5>
- Chen, Q., Wu, Y., Wang, X., Jiang, Z. L., Zhang, W., Liu, Y., & Alazab, M. (2023). A Generic Cryptographic Deep-Learning Inference Platform for Remote Sensing Scenes. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16, 3309-3321. doi:10.1109/JSTARS.2023.3260867
- Chenyi Chen, Ari Seff, Alain Kornhauser, Jianxiong Xiao(2015). Deepdriving: Learning affordance for direct perception in autonomous driving . *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2722-2730. doi: 10.1109/ICCV.2015.312
- Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R. & Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3213-3223). doi: 10.1109/CVPR.2016.350
- Dorrani, Z. (2024). Traffic Scene Analysis and Classification using Deep Learning. *International Journal of Engineering*, 37(3), 496-502. doi: 10.5829/ije.2024.37.03c.15
- Dutta, R., Aryal, J., Das, A., & Kirkpatrick, J. B. (2013). Deep cognitive imaging systems enable estimation of continental-scale fire incidence from climate data. *Scientific reports*, 3(1), 3188. doi: 10.1038/srep03188
- Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88, 303-338. doi: 10.1007/s11263-009-0275-4

- Ganashree, K. C. G., Hemavathy, R., & Anala, M. R. (2024). Land scene classification from remote sensing images using improved artificial bee colony optimization algorithm. *International Journal of Electrical & Computer Engineering* (2088-8708), 14(1).doi: 10.11591/ijece.v14i1.pp347-357
- Gould, S., Fulton, R., & Koller, D. (2009, September). Decomposing a scene into geometric and semantically consistent regions. In *2009 IEEE 12th international conference on computer vision* (pp. 1-8). IEEE. doi: 10.1109/ICCV.2009.5459207
- Hao, Z., Zhan, H., Zhang, C., Pei, N., Sun, B., He, J. & Wang, C. (2022). Assessing the effect of human activities on biophony in urban forests using an automated acoustic scene classification model. *Ecological Indicators*, 144, 109437. <https://doi.org/10.1016/j.ecolind.2022.109437>
- Hasan, N. W., Saudi, A. S., Khalil, M. I., & Abbas, H. M. (2022). A genetic algorithm approach to automate architecture design for acoustic scene classification. *IEEE Transactions on Evolutionary Computation*, 27(2), 222-236. <https://doi.org/10.1109/TEVC.2022.3185543>
- Hussain Raza, S., Grundmann, M., & Essa, I. (2013). Geometric context from videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3081-3088). <https://doi.org/10.48550/arXiv.1510.07320>
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 25. <http://dx.doi.org/10.1145/3065386>
- Kumar, S., & Setty, S. L. N. (2023). UFS-LSTM: unsupervised feature selection with long short-term memory network for remote sensing scene classification. *Evolutionary Intelligence*, 16(1), 299-315. doi:10.1007/s12065-021-00660-4
- Laulkar, C. A., & Kulkarni, P. J. (2020). Semantic rules-based Classification of outdoor natural scene images. In *Computing in Engineering and Technology: Proceedings of ICCET 2019* (pp. 543-555). Springer Singapore. doi: 10.1007/978-981-13-8715-9_52
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, 521(7553), 436-444. <https://doi.org/10.1038/nature14539>
- Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D. & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13* (pp. 740-755). Springer International Publishing. doi: 10.1007/978-3-319-10602-1_48
- Liu, C., Yuen, J., & Torralba, A. (2009, June). Nonparametric scene parsing: Label transfer via dense scene alignment. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1972-1979). IEEE. <http://dx.doi.org/10.1109/CVPRW.2009.5206536>
- Liu, Z., Dong, A., Yu, J., Han, Y., Zhou, Y., & Zhao, K. (2022). Scene classification for remote sensing images with self-attention augmented CNN. *IET Image Processing*, 16(11), 3085-3096. doi: 10.1049/ipr2.12540
- Mottaghi, R., Chen, X., Liu, X., Cho, N. G., Lee, S. W., Fidler, S. & Yuille, A. (2014). The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 891-898). doi: 10.1109/CVPR.2014.119
- Jayanth, H. N. (2020). Natural Scene Image Classification using CNN in *International Journal of Advance Research, Ideas and Innovations in Technology* (pp. 79-86)
- Nadeem, U., Shah, S.A.A., Soheli, F., Togneri, R., Bennamoun, M. (2019). Deep Learning for Scene Understanding. In: Balas, V., Roy, S., Sharma, D., Samui, P. (eds) *Handbook of Deep Learning Applications. Smart Innovation, Systems and Technologies*, vol 136. Springer, Cham. https://doi.org/10.1007/978-3-030-11479-4_2
- Naseer, A., Al Mudawi, N., Abdelhaq, M., Alonazi, M., Alazaib, A., Algarni, A., & Jalal, A. (2024). CNN-based Object Detection via Segmentation capabilities in Outdoor Natural Scenes. *IEEE Access*. doi: 10.1109/ACCESS.2024.3413848
- Neupane, B., Horanont, T., & Aryal, J. (2021). Deep Learning-Based Semantic Segmentation of Urban Features in Satellite Images: A Review and Meta-Analysis. *Remote Sensing*, 13(4), 808. <https://doi.org/10.3390/rs13040808>
- Nweke, H. F., Teh, Y. W., Al-Garadi, M. A., & Alo, U. R. (2018). Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges. *Expert Systems with Applications*, 105, 233-261. <https://doi.org/10.1016/j.eswa.2018.03.056>
- Oktavian, G. M., & Santoso, H. (2024, April). Improving accuracy of multi-class image classification using ensemble learning of convolutional neural networks. In *AIP Conference Proceedings* (Vol. 3070, No. 1). AIP Publishing. <https://doi.org/10.1063/5.0207151>
- Patel, H., & Mewada, H. (2018). Analysis of machine learning based scene classification algorithms and quantitative evaluation. *International Journal of Applied Engineering Research*, 13(10), 7811-7819.

- Patel, T. A., Dabhi, V. K., & Prajapati, H. B. (2020, March). Survey on scene classification techniques. In *2020 6th international conference on advanced computing and communication systems (ICACCS)* (pp. 452-458). IEEE. <https://doi.org/10.1109/ICACCS48705.2020.9074460>
- Pereira, R., Barros, T., Garrote, L., Lopes, A., & Nunes, U. J. (2024). A deep learning-based global and segmentation-based semantic feature fusion approach for indoor scene classification. *Pattern Recognition Letters*, 179, 24-30. <https://doi.org/10.1016/j.patrec.2024.01.022>
- Pfeiffer, D., Gehrig, S., & Schneider, N. (2013). Exploiting the power of stereo confidences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 297-304). doi: 10.1109/CVPR.2013.43.
- Quattoni, A., & Torralba, A. (2009, June). Recognizing indoor scenes. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 413-420). IEEE. doi: 10.1109/CVPR.2009.5206537
- Raja, R., M, S., Roomi, M. and Dharmalakshmi, D. "Classification and retrieval of natural scenes," *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*, Tiruchengode, India, 2013, pp.1-8. doi: 10.1109/ICCCNT.2013.6726534.
- Ran, Y., Xu, X., Luo, M., Yang, J., & Chen, Z. (2023). Scene Classification Method Based on Multi-Scale Convolutional Neural Network with Long Short-Term Memory and Whale Optimization Algorithm. *Remote Sensing*, 16(1), 174. <https://doi.org/10.3390/rs16010174>
- Russell, B. C., Torralba, A., Murphy, K. P., & Freeman, W. T. (2008). LabelMe: a database and web-based tool for image annotation. *International journal of computer vision*, 77, 157-173. doi:10.1007/s11263-007-0090-8.
- Salakhutdinov, R., & Larochelle, H. (2010, March). Efficient learning of deep Boltzmann machines. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (pp. 693-700). JMLR Workshop and Conference Proceedings.
- Salem, M. A. M., Atef, A., Salah, A., & Shams, M. (2017). Recent survey on medical image segmentation. In *Handbook of Research on Machine Learning Innovations and Trends* (pp. 424-464). IGI global. doi: 10.4018/978-1-5225-2229-4.ch019
- Shadman Roodposhti, M., Aryal, J., Lucieer, A., & Bryan, B. A. (2019). Uncertainty assessment of hyperspectral image classification: Deep learning vs. random forest. *Entropy*, 21(1), 78. doi:10.3390/e21010078
- Shafaey, M. A., Salem, M. A. M., Al-Berry, M. N., Ebied, H. M., El-Dahshan, E. A., & Tolba, M. F. (2020, March). Hyperspectral image classification using deep learning technique. In *The International Conference on Artificial Intelligence and Computer Vision* (pp. 334-342). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-44289-7_31
- Shalaby, M. M., Mohammed, A., Salem, M., Khamis, A., & Melgani, F. (2014, December). Geometric model for vision-based door detection. In *2014 9th International Conference on Computer Engineering & Systems (ICCES)* (pp. 41-46). IEEE. doi:10.1109/ICCES.2014.7030925
- Shao, Y. F., Ma, X. X., Ma, Y., & Zhang, W. Q. (2024). Deep semantic learning for acoustic scene classification. *EURASIP Journal on Audio, Speech, and Music Processing*, 2024(1), <https://doi.org/10.1186/s13636-023-00323-5>
- Shen, D., Ji, Y., Li, P., Wang, Y., & Lin, D. (2020). Ranet: Region attention network for semantic segmentation. *Advances in Neural Information Processing Systems*, 33, 13927-13938. <http://hdl.handle.net/10397/105505>
- Shi, J., Zhu, H., Li, Y., Li, Y., & Du, S. (2022). (Retracted) Scene Classification Using Deep Networks Combined with Visual Attention. *Journal of Sensors*, 2022(1), 7191537. <https://doi.org/10.1155/2023/9818617>
- Silberman, N., Hoiem, D., Kohli, P., & Fergus, R. (2012). Indoor segmentation and support inference from rgb-d images. In *Computer Vision—ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12* (pp. 746-760). Springer Berlin Heidelberg. Doi: 10.1007/978-3-642-33715-4_54
- Simonyan, K. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Sitaula, C., KC, S., & Aryal, J. (2024). Enhanced multi-level features for very high resolution remote sensing scene classification. *Neural Computing and Applications*, 36(13), 7071-7083. doi: 10.1007/s00521-024-09446-y
- Sitaula, C., Shahi, T. B., Marzbanrad, F., & Aryal, J. (2024). Recent advances in scene image representation and classification. *Multimedia Tools and Applications*, 83(3), 9251-9278. doi:10.1007/s11042-023-15005-9.
- Sitaula, C., Xiang, Y., Aryal, S., Lu, X. (2019). Unsupervised Deep Features for Privacy Image Classification. In: Lee, C., Su, Z., Sugimoto, A. (eds) *Image and Video Technology. PSIVT 2019. Lecture Notes in Computer Science()*, vol 11854. Springer, Cham. https://doi.org/10.1007/978-3-030-34879-3_31
- Sitaula, C., Xiang, Y., Zhang, Y., Lu, X., & Aryal, S. (2019). Indoor image representation by high-level semantic features. *IEEE Access*, 7, 84967-84979. doi: 10.1109/ACCESS.2019.2925002

- Sivasubramanian, A., VR, P., V, S., & Ravi, V. (2024). Transformer based ensemble deep learning approach for remote sensing natural scene classification. *International Journal of Remote Sensing*, 45(10), 3289-3309. <https://doi.org/10.1080/01431161.2024.2343141>
- Song, D., Liu, H., Dong, Q., Bian, Z., Wu, H., & Lei, Y. (2018). Digital, rapid, accurate, and label-free enumeration of viable microorganisms enabled by custom-built on-glass-slide culturing device and microscopic scanning. *Sensors*, 18(11), 3700. doi:10.3390/s18113700
- Soudy, M., Afify, Y. M., & Badr, N. (2022). GenericConv: A Generic Model for Image Scene Classification Using Few-Shot Learning. *Information*, 13(7), 315. <https://doi.org/10.3390/info13070315>
- Tang, Q., Liu, F., Zhang, T., Jiang, J., Zhang, Y., Zhu, B., & Tang, X. (2022). Compensating for local ambiguity with encoder-decoder in urban scene segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 23(10), 19224-19235. doi: 10.1109/TITS.2022.3157128
- Torralba, A., Walther, D. B., Chai, B., Caddigan, E., Fei-Fei, L., & Beck, D. M. (2013). Good exemplars of natural scene categories elicit clearer patterns than bad exemplars but not greater BOLD activity. *PloS one*, 8(3), e58594. doi: 10.1371/journal.pone.0058594
- Xiao, J., Owens, A., & Torralba, A. (2013). Sun3d: A database of big spaces reconstructed using sfm and object labels. In *Proceedings of the IEEE international conference on computer vision* (pp. 1625-1632).
- Yang, L. I., Zhang, L. I., Zi, W. A. N. G., Kun, W. A. N. G., & Qifeng, Y. U. (2024). Semi-supervised remote sensing image scene classification with prototype-based consistency. *Chinese Journal of Aeronautics*, 37(2), 459-470. <https://doi.org/10.1016/j.cja.2023.12.012>
- Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., ... & Darrell, T. (2020). Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2636-2645). doi:10.1109/CVPR42600.2020.00271
- Zeng, D., Liao, M., Tavakolian, M., Guo, Y., Zhou, B., Hu, D., ... & Liu, L. (2021). Deep learning for scene classification: A survey. *arXiv preprint arXiv:2101.10531*.
- Zhang, Q., Liu, Y., Gong, C., Chen, Y., & Yu, H. (2020). Applications of deep learning for dense scenes analysis in agriculture: A review. *Sensors*, 20(5), 1520. <https://doi.org/10.3390/s20051520>
- Zhao, R., Wang, D., Yan, R., Mao, K., Shen, F., & Wang, J. (2017). Machine health monitoring using local feature-based gated recurrent unit networks. *IEEE Transactions on Industrial Electronics*, 65(2), 1539-1548. doi:10.1109/TIE.2017.2733438
- Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., & Torralba, A. (2017). Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 633-641). doi: 10.1109/CVPR.2017.70
- Zhou, Z., Li, S., Wu, W., Guo, W., Li, X., Xia, G., & Zhao, Z. (2021). NaSC-TG2: Natural scene classification with Tiangong-2 remotely sensed imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14, 3228-3242. doi: 10.1109/JSTARS.2021.3063096

Priti

Deenbandhu Chhotu Ram University of
Science and Technology, Murthal,
India
pritiagrover416@gmail.com
ORCID 0009-0007-0526-6886

Dinesh Singh

Deenbandhu Chhotu Ram University of
Science and Technology, Murthal,
India
dineshsingh.cse@dcrustm.org
ORCID 0000-0002-7614-5789

Savita Ahlawat

Maharaja Surajmal Institute of
Technology, New Delhi,
India
amarjit.malhotra@nsut.ac.in
ORCID 0000-0002-1191-9173
