SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V11I9P102

Volume 11 Issue 9, 11-21, September 2024

Original Article

CNN Architectures for Image Classification: A Comparative Study Using ResNet50V2, ResNet152V2, InceptionV3, Xception, and MobileNetV2

Nitin Duklan[1], Sachin Kumar[1], Himani Maheshwari[2], Rajesh Singh[3], Sameer Dev Sharma[4], Siddharth Swami[5]

[1]*Uttaranchal Institute of Technology, Uttaranchal University, Uttarakhand, India.*
[2]*Department of CSE, Graphic Era Hill University, Uttarakhand, India.*
[3]*Division of Research & Innovation, Uttaranchal University, Uttarakhand, India.*
[4]*Uttaranchal School of Computing Science, Uttaranchal University, Uttarakhand, India.*
[5]*School of Environment and Natural Resources, Doon University, Dehradun, Uttarakhand, India.*

[5]*Corresponding Author : siddharthswami3@gmail.com*

Received: 21 June 2024 Revised: 03 August 2024 Accepted: 26 August 2024 Published: 30 September 2024

Abstract *- Image processing techniques have been used for picture categorization in several domains over the last year, including education, research, railways, and other sectors. The CNN (Convolutional Neural Network) is often regarded as the most potent method for picture categorization. This study included five renowned image-processing algorithms using the CNN architecture: RestNet50V2, RestNet152V2, Xception, Inceptionv3, and MobileNetV2. We assessed the classification of the Uttaranchal University, Dehradun dataset, which has 20 distinct department photos for classification. After a certain iteration, our primary goal is to achieve the best model accuracy possible using the available hardware. To evaluate performance, we used other measures such as accuracy, recall, and F1-score. The investigation demonstrated the exceptional precision of all five algorithms: RestNet50V2 (98.88), RestNet152V2 (99.10), Xception (99.17), Inceptionv3 (99.2), and MobileNetV2 (93.71). The Xception method was chosen for data training, testing, and validation because of its superior accuracy. Hardware resources, memory capacity, and data diversity are also considered while assessing algorithm pros and cons. This research sheds light on the CNN model's performance and helps companies and universities choose better photo classification algorithms. This research has also advanced machine learning and deep learning algorithms, as well as their practical application in real-world situations.*

Keywords *- CNN, Inceptionv3, MobileNetV2, RestNet50V2, RestNet152V2, Xception.*

1. Introduction

Deep learning, especially CNNs, has excelled in picture identification and classification in recent years. The identification and categorization of objects in photographs have consequences for education, healthcare, and security. This work uses state-of-the-art CNN architectures (ResNet50V2, ResNet152V2, InceptionV3, Xception, and MobileNetV2) to identify departments at Uttaranchal University, Dehradun. University department identity is critical for administrative, student, and resource allocation objectives.

Traditional department identification techniques require a laborious physical examination or barcode scanning, which might be inaccurate. Deep learning techniques can be used to automate this identification procedure and will improve overall efficiency and reliability. The need for an automated department identification system that could analyze enormous amounts of photographs accurately and quickly encouraged our study.

To achieve this, approximately 3000 photographs were gathered from different departments at different locations of the Uttaranchal University, Dehradun campus, and thereafter annotated, individual pictures with the department name, which will assist Convolutional Neural network model supervised learning. For Image categorization, accuracy, efficiency, and scalability are the important criteria for the selection of CNN architectures, ResNet50V2, ResNet152V2, Xception, MobileNetV2, and Inception designs architectures have excellent accuracy, efficiency, and scalability on picture categorization performance on standard datasets. This study focuses on creating a reliable and accurate department identification system that will integrate with the university's infrastructure and might improve Uttaranchal University's administrative operations, such as campus security and student experience.

Recently, deep learning models like ResNet50V2, ResNet152V2, InceptionV3, Xception, and MobileNetV2 have gained popularity for picture categorization and object identification to solve picture recognition problems. This study examines various designs that were used for comparative tasks and discusses their performance, benefits, and drawbacks. He et al. [1]invented Residual Networks (ResNet) and proposed a residual learning framework to ease the training of deeper networks using skip connections, solving the exploding/vanishing gradient issue and improving accuracy and convergence speed. Image classification experiments have shown ResNet variations like ResNet50V2 and ResNet152V2 to be successful [2][3]. The design proposed by Chollet [2] improved performance measures of top-picture categorization while reducing computational complexity and memory footprint. This design proposed extending factorised convolutions and replacing the existing classical convolutions with depthwise separable convolutions. Xception outperforms alternative designs on certain datasets.

The Inception architecture, developed by Szegedy et al. [4], uses mixed convolutional filters of various sizes to capture spatial hierarchy in images effectively. The processing performance and the ability of InceptionV3 architecture to effectively balance the model's complexity made this architecture popular for image recognition[5]. MobileNetV2 mobile architecture by Sandler et al. [3], for image categorization and object identification specifically optimized for handheld/ embedded/ mobiles with minimum processing resources, has excellent accuracy with low processing cost by using inverted residuals and linear bottlenecks. Many studies have examined the performance of different designs on typical picture classification benchmarks, such as ImageNet. Tan et al. [6] evaluated ResNet, Inception, Xception, and MobileNetV2 on many datasets, emphasizing accuracy, speed, and model size trade-offs.

Howard et al. [7] used mobile devices to evaluate designs, emphasizing model efficiency for real-time applications. Campus Management: Although there is less study on department identification tasks in universities, general image classification concepts and methods may be applied to this sector. Deep learning models may be used in campus management applications after studies like [8] and [9] have shown their versatility. Due to their speed, efficiency, and scalability, ResNet50V2, ResNet152V2, InceptionV3, Xception, and MobileNetV2 are the top image classification architectures. We want to use literature to assess these architectures for department identification at Uttaranchal University and enhance deep learning research in campus administration applications. The study showcases that deep learning for image classification has excellent accuracy across different CNNs architectures here in this study different metrices like scalability, latency, efficiency, and computing for real world model deployment were examined. To improve models' resilience advanced preprocessing techniques and data augmentation methods are used. The study compares notable CNN architectures and presents a sequential method for dataset preparation, training on the dataset and assessment guidelines that can help in model selection. Addressing future aspects that can enhance image processing and developing a robust and accurate model includes the use of Transfer learning, attention processes, domain adaptability, ensemble learning, and real-world deployment.

2. Literature Review

The purpose of the literature review is to provide a thorough and inclusive summary of the latest progress in deep learning structures and methodologies for computer vision assignments. An analysis and synthesis of the following influential works is conducted to determine the main contributions and trends in the subject. Over the last several decades, image processing has advanced greatly. This review breaks down image processing algorithms, assesses their efficacy, and illuminates new paradigms. We review traditional and deep learning techniques to offer a complete picture. Our study includes image processing studies, conference papers, and patents. We examine classics and innovations. Convolutional Neural Network (CNN) techniques like ResNet50V2, ResNet152V2, Xception, and MobileNetV2 are nowadays used in many fields for image processing. When it comes to computer vision applications, ConvNet/CNNs are outperforming conventional methods for image processing approaches [10] [11] and proposed an accurate and precise surveillance system for crime mitigation and disease detection in rice using a depthwise separable CNN based Xception model. Although significant studies discuss that different CNN designs have distinct performance characteristics and outperform certain classical methods, in several tasks, Xception outperforms VGG16 and EfficientNetV2.[10] CNNs, when merged along with LSTM and RNN networks, increase the overall performance[12] [13].

Overall, studies reflect that when CNNs are integrated with diverse neural networks, they may also significantly enhance image processing [12] [13]. CNN architectures like Xception are frequently on top, and ResNet outperforms typical image processing methods and significantly impacts image processing procedures. It has been extensively studied that image processing is used to automate image processing and increase picture quality in computer vision [14]. Analyzing and comprehending picture materials in different transdisciplinary research uses neuro-computing Deep Learning (DL).[15] These advancements have also improved transmission line identification and damage detection [16]. Computer vision has made significant progress, but there are gaps and problems. Some of the issues remain unresolved, like Key image processing, which requires more exploration despite advancements in Deep learning and Computer vision. Therefore, to fill the research gaps and move forward in this area, the current tier provides a potent deck for conducting a

scientific study. Data pre-processing is an important aspect of improving analytical performances, accuracy, and resilience and enhancing machine learning models in image processing. It includes data preparation [1], augmentation and integration. These methods remove any noise and decrease CT slice intensity. Image processing includes data preparation and data augmentation, and The data preparation process transforms raw data through various methods like thresholding, morphological operation, histogram equalization or contrast enhancement, etc. In contrast, data augmentation methods through scaling, rotation, cropping, flipping and translating create a more representative and heterogenous dataset. The data augmentation method improves unbalanced class performance and prevents overfitting of the neural network, thus generating better recognition results. Recent studies discuss and highlight the prerequisite of data augmentation.

In this study, many image classification and data augmentation approaches are evaluated, starting with rotating, cropping, zooming, and histogram-based image alterations and then moving forward to style transfer and generative adversarial networks. Our innovative image style transfer-based data augmentation method generates high-quality images that blend basic image information with additional styles. Our image classification system seeks to increase training efficiency and diagnosis accuracy. Finally, data pretreatment and augmentation strengthen machine learning models. Overfitting, insufficient training data, and unequal class composition are all handled. The year 2015 saw the introduction of ResNet [17], and it quickly gained popularity for image classification. ResNet design uses residual blocks to develop a residual mapping between input and output feature maps, improving network accuracy. He et al. put out the concept of deep residual learning as a solution to the difficulty of training neural networks with many layers. Their proposed model displayed exceptional performance along with facilitating efficient gradient propagation over skip links. Chollet [2] released Xception, a modified Inception, in 2016.

Picture categorization in CNNs, such as Xception, uses depth-wise separable convolutions. This deep learning structure reduces computational complexities and enhances performance while reducing computational cost, making it suitable compared to traditional approaches. Sandler et al.[3] proposed MobileNetV2, which is designed for devices with limited processing capabilities. This architecture uses parallel depthwise separable CNN and, therefore, minimizes memory and processing resources and reduces the bandwidth requirement. They also used inverted residuals and linear bottlenecks to enhance the efficiency and precision of mobile image recognition tasks. In their study, Rezatofighi et al. [5], to precisely assess object localization, proposed a Generalized Intersection over Union (GIoU) metric and loss function. This metric results in recognising objects more accurately and enhances the accuracy of the bounding box regression model by overcoming the drawbacks of existing metrics. A new technique called deep fusion was introduced by Xie et al. [18], in which multiscale convolutional features were used to evaluate image quality. This multimedia augmentation and content analysis methodology offers crucial perspectives of image processing architectures. An efficient facial expression detection mechanism for handheld devices was again an unexplored area till Jiang et al.[19]system based on MobileNetV2 effectively handles the issues associated with real-time facial expression analysis and identification for encouraging outcomes for practical uses. Huang et al. [20] created DenseNet to overcome fading gradients and encourage feature reuse. To enhance picture identification, [4]Szegedy et al. explored how residual connections affect deep neural network training. Tan and Le (2019) [21] developed EfficientNet to scale convolutional neural networks and increase performance while lowering parameters.

Howard et al. [6] MobileNets aid vision applications with their low latency and high accuracy. MobileNets are popular mobile apps because of their speed and precision. SSD was introduced by [8] for object identification. Real-time identification is done using a sophisticated neural network. Due to these developments, image classification speed and benchmarks have improved. EfficientNet outperformed prior models on ImageNet with fewer parameters SSD accurately predicted item bounding boxes and class probabilities across sizes with exceptional speed. Simonyan and Zisserman [9] developed sophisticated convolutional networks for large-scale image recognition. This research laid the groundwork for deep architectures like VGGNet. Deep learning achieved the best photo classification performance in their study. Lin et al. [22] presented Network in Network (NiN), which optimizes deep neural networks by including tiny neural networks.

NiN improves computer vision performance by acquiring features at several abstract layers. Zeiler and Fergus [23] pioneered convolutional network visualization for maximum activation and to elucidate deep neural network they use visualization methods like deconvolutional networks. To overcome resolution variations problems during training and testing in deep learning model Touvron et al. [24] introduced a resolution independent analysis procedure that will ensure reliability by incorporating fail model evaluation across different resolutions and elevate computer vision research reproducibility. For analyzing videos Wang and Gupta proposed a unique approach for representing movies using space-time area graphs which provides a comprehension and exceptional video analysis. Their approach accurately represents temporal connections among video frames as a result competence in action identification and video segmentation is achieved. For handwritten digit recognition Ciresan et al. [25] suggested MNIST digit recognition system that excels deep, large, and simple neural networks for the same. For the compression and transmission of models projecting it to deploy on machines with limited resources Hinton et al. [26] explored the concept of knowledge

distillation that trains a relatively smaller model to imitate the behavioral concepts of larger models. To segment biological images Ronneberger, Fischer, and Brox [27] developed U-Net a symmetrical encoder-decoder that routes and skips connections and enables excellent image segmentation for biological images. With limited training data image categorization and classification in a big challenge and in 2017 [28] developed a model AlexNet with a deep convolutional neural network this idea revolutionized categorization and classification of images on ImageNet so that even with limited training dataset images can be accurately segmented.

Sun et al. [29], for collecting hierarchical info at different levels while increasing visual task discrimination, developed a multi-scale order with less pooling of deep convolutional activation features that improve robustness while presenting images by aggregate features. Selective search is essential for object recognition and segmentation, allowing accurate and fast object location. Uijlings et al.[30]developed a selective search technique for object detection that provides a variety of region suggestions for input data while improving accuracy and variety. Deep learning has opened various possibilities in video compression and content suggestion[31]. Karpathy et al., using Deep CNNs, categorize and classify the film data on a very large scale in an efficient manner. Image classification models are used to map and understand models' behavior silently, and deep brain network functions are used (Simonyan, Vedaldi, and Zisserman[32]. TransGaN, an adversarial generative network created by Li, Arnab, and Torr [33]for multimodal zero-shot learning, provides a diverse set of various new classes that use adversarial training to generate cross-modal representations allowing zero hot learning in multimodal environments. Another model that improves accuracy and speed and detection for object detection framework is Fast R-CNN, developed by Girshick [34], which is ideal for real-time applications. To improve tracking accuracy in different situations, Nam and Han [35]trained context-adaptive CNN across visual domains as a method that adopts feature representations.

Gulzar [36] A deep learning model was developed based on MobileNetV2 architecture. In the proposed model, five different layers were added after removing the classification layer present in the MobileNetV2 architecture to improve the efficiency and accuracy of the model. [37] Their research analyzed the performance of the MobileNetV2 model for image classification and achieved higher accuracy. [38] Aims to compare the performance of ResNet50V2 and MobileNetV2 in architectural style classification and has justified that ResNet50V2 showcases high accuracy and stability when compared with MobileNetV2, which is stable with fluctuations

To summarize, the literature study emphasizes notable progress in deep learning structures and methodologies for computer vision assignments. These studies have made significant contributions to the advancement of efficient and effective models for image recognition, object identification, and facial expression analysis, hence facilitating future studies on the subject.

3. Materials and Methods

3.1. Data Collection

The study report on Uttaranchal University, Dehradun, aimed to get a thorough comprehension of the campus's spatial utilization, traffic patterns, and facility accessibility. Researchers used geo-tagging technology to accurately document important information such as pedestrian movement, facility use, and spatial arrangements. Subsequently, all this data was used to do geographic mapping and analysis.

The researchers focused primarily on the Administrative Block and conducted thorough monitoring of administrative activity, visitor movement, and the use of office space in the building. A complete evaluation was conducted on the study habits of students, the accessibility of library resources, and the capacity of the Central Library. The objective of this research was to examine the level of academic rigor, utilization of laboratory facilities, and the dynamics of professor-student interactions in engineering courses, namely the University Aerospace Department, University Mechanical Workshop, and University Civil Engineering. To correlate sports involvement and recreational activities, this study was focused on specialized places like the University Gym and University grounds.

Further, to ensure the reliability and validity of the dataset, we used technical breakthroughs such as GPS-equipped gadgets and data recording software. The glocalization data of the University campus is comprehensively gathered and mapped accurately. For the utilization of campus resources, an examination was conducted on different venues like parking, cafeterias, food courts, mess, and sports complexes. We collected data from Uttaranchal University, which includes images from the 20 different departments on the university campus.

Each department provides a distinct class label with geo-location for better identification of the department, which resulting this study being a multi-class classification problem. The dataset contains various types of images, including interior and exterior parts of university buildings, facilities, and infrastructure. The original dataset contains a total of 2882 images of different departments of the university. Image preprocessing and various data augmentation techniques enhanced the input images in a total of 26,782 images for training and evaluation. Data collection was carried out using ethical principles to guarantee the protection of information privacy and confidentiality. We acquired permission and de-identified sensitive data to protect anonymity.

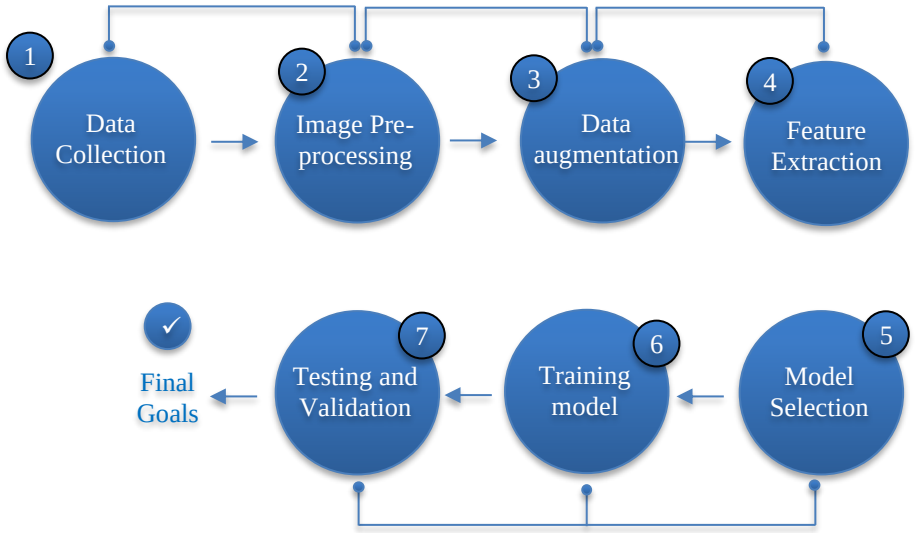

Fig. 1 The methodological procedure of this research

Figure 1 illustrates the technique that was followed in this experiment, which included collecting images, preprocessing those images, augmentation of those images, feature engineering, model selection, and training and testing the model. The twenty classes of Uttaranchal University are shown in Figure 2, which displays the statistics for all the departments. The y-axis displays the number of photos belonging to each class, while the x-axis represents the class name. Figures 3 and 4 display the example photos of all 20 classes together with their corresponding geo-locations.

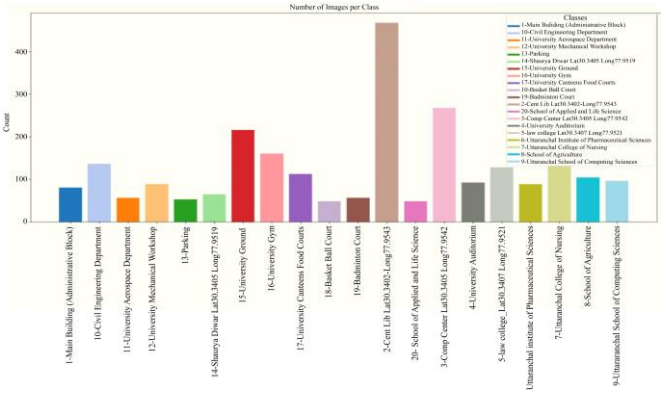

Fig. 2 The whole dataset of the 20 classes at Uttaranchal University

Fig. 3 Sample images of all 20 classes with geo-location

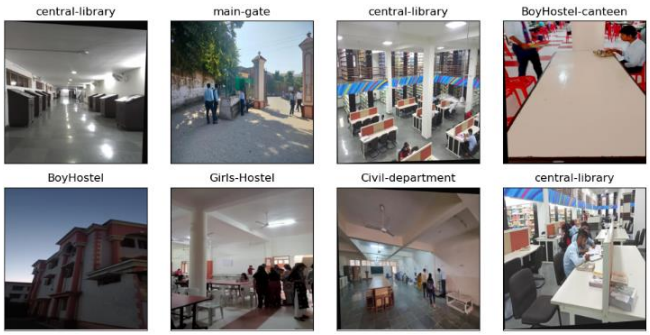

Fig.4 Sample images of some classes

The study report on Uttaranchal University used a comprehensive and meticulous data-gathering approach by integrating physical observations with advanced technical equipment.

The information provided may provide valuable insights into various departments and infrastructures, which can assist in analyzing campus dynamics, optimizing resource allocation, and guiding strategic decision-making for the university's future growth.

3.2. Data Preprocessing

Data preprocessing is a crucial stage in preparing data for analysis and visualization, as explained in the research paper. The images from Uttaranchal University were improved by utilizing several methods to ensure their consistency, use, and excellence. The first step, known as Auto-Orient, automatically adjusts the rotation of the image to minimize distortion and discrepancies. The Static Crop technique removes unnecessary areas and boundaries by focusing just on the focused region of concern. The photographs undergo refinement in the horizontal and vertical region preprocessing stages by being separated into their respective regions and then retaining just the central area.

Several methods, such as cropping, automatic contrast adjustment, and dynamic cropping, along with other methods, were used to improve the standard and usability and prepare them for machine learning models. Photographs are consistently reduced to 640 X 640 pixels dimensions to improve performance across different platforms. To reduce the complexity of image computing and improve understandability and analytical capabilities, images are converted to greyscale. Further, to compare multiple images simultaneously, a tile preparation step was followed where images were arranged in a grid format. Figure 5 depicts various image preprocessing steps.

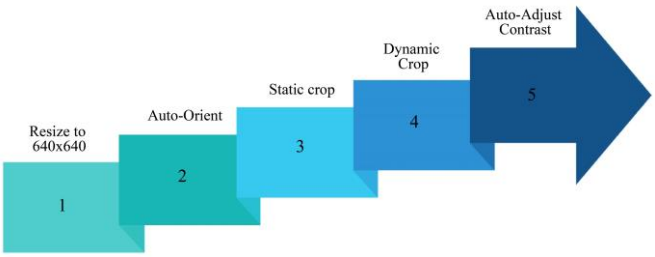

Fig. 5 Image preprocessing steps

Ultimately, preprocessing methods like Grayscale conversion, Tile organization, Horizontal/Vertical region cropping, auto orientation, Resize, stretch and static crop are used in this work to facilitate the preparation of a standardized dataset to enhance the validity and reliability of the research findings. These parameters enhance the validity and reliability of the research findings.

3.3 Data Augmentation

The study results provide an entire data augmentation approach for image processing applications. Adopting this method enhances the robustness and versatility of deep neural network models. The augmentation approaches include many geometric alterations that impact the picture alignment. The modifications comprise a ninety-degree horizontal inversion, a clockwise rotation, and an anticlockwise rotation. Cropping strategies that approximate size and cropping modifications include tactics like Crop 0% Minimum Zoom and Crop 20% Maximum Zoom. Horizontal and vertical shears of ±10 degrees each, as well as rotations ranging from -15 degrees to +15 degrees, have the potential to induce visual distortion. Moreover, these errors boost the model's resilience to changes in the shape and position of objects, accurately mimicking the mistakes that occur in real-life scenarios.

To increase the model's power to handle typical picture distortions and its ability to generalize, two further improvements are implemented: Blur up to 2.5 pixels and noise up to 0.1% of pixels to simulate these aberrations. Figure 6 illustrates the many data augmentation techniques, such as flipping, 90-degree rotation, hue adjustment, saturation adjustment, brightness adjustment, exposure adjustment, cropping, rotating, shearing, and adding noise.

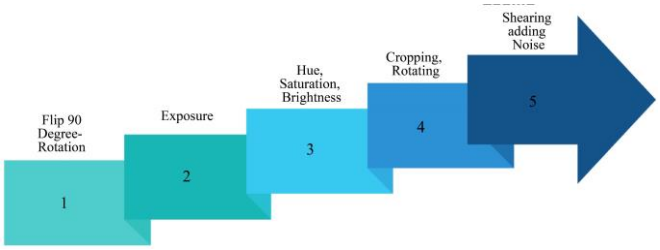

Fig. 6 The sequential process of data augmentation

The larger dataset of 26,782 images, which underwent augmentation and preprocessing, provided a more complete view of the distribution of the underlying data. This research aims to increase the model's adaptability to new data sets and real-world outcomes. Create a robust dataset to train image classification algorithms.

3.4 Feature Extraction Methods

In this investigation, we used ResNet50V2, ResNet152V2, InceptionV3, Xception, and MobileNetV2 deep learning frameworks, together with a range of feature extraction methods and machine learning models. Here, we provide an overview of several techniques and models:

3.4.1. Techniques for Extracting Features

1. Convolutional Neural Networks (CNNs): CNNs are deep learning models that are specifically designed for image processing tasks. We used pre-trained Convolutional Neural Network (CNN) architectures such as ResNet, VGG, or MobileNet to extract hierarchical features from the input pictures. CNNs use convolutional layers and pooling operations to acquire significant visual representations autonomously.
2. HOG, a traditional feature extraction approach, is often employed in computer vision problems. Compute the gradient orientation distribution in local picture areas. HOG descriptors were used to collect shape and edge information from input pictures when CNN-based features were not possible.
3. Scale-Invariant Feature Transform (SIFT): A classic feature extraction approach for object detection and picture matching. It finds picture key points and computes scale-, rotation-, and illumination-invariant descriptors. SIFT descriptors were used to extract unique features from input photos, particularly where transformation resilience is important.

This research aims to improve picture classification by using pre-trained Convolutional Neural Networks (CNNs). A pre-trained model saves time and resources via the transfer of learning method. To reduce overfitting and underfitting and ensure the CNN model performs the specific task in an optimized way, fine-tuning the model helps to improve the weight for better results. Ensemble learning methods, such as averaging or stacking, address the problem of overfitting and

use the collective knowledge of several models to improve the accuracy and generalizability of predictions.

Feature extraction approaches and machine learning algorithms are used to extract unique characteristics from input photos and build prediction models that can precisely categorise or analyze them based on research goals.
Transfer learning is used to customize the information generated from existing models for different tasks and minimize the need for training data. Also, to improve the reliability and performance of a model, Ensemble learning is used.

To address the challenges in image recognition, this approach uses deep learning architectures such as ResNet50V2, ResNet152V2, InceptionV3, Xception, and MobileNetV2. This approach not only boosts the system efficiency but also improves dependability and durability in real-time problems with reduced resource usage and computational costs.

Enhancements in this model are incorporated to gain the capabilities and flexibility like ResNet50V2, ResNet152V2, InceptionV3, Xception, and MobileNetV2 by increasing the number of epochs, and the cyclic learning rate approach gives good results after a number of epochs leading to better handling of over and underfitting. The weight decay and dropouts avoid overfitting and thus regularize procedures.

3.5. Evaluation Metrics

The standard evaluation metrics in any machine learning model or experiment are accuracy, precision, recall and F-1 Score. Accuracy is an important factor in determining a model's performance. If a model is accurate, it is also a reliable model; here, accuracy is the ratio of the number of correctly classified instances to the total number of occurrences.

Precision primarily focuses on the positive indicators, which is the ratio of true positives (correctly predicted positive values) out of the total number of anticipated positives, avoiding false positive values. Recall or sensitivity emphasizes recording all positive instances. It is a metric that measures the ration of predicted positives to the total real positive instances.

F1-score is an evaluation metric that uses or combines the precision and recall scores rather than accuracy. It is basically a harmonic mean of recall and precision that considers false positives and false negatives.

4. Hardware and Software Requirements

The study was conducted using supervised learning/computing environments along with deep learning approaches. The hardware system used for the computation has a GPU with 64 GB memory and an NVIDIA GeForce GTX 1070 Ti, which is sufficient for neural network training. Several software programs like Anaconda, Python, Jupyter, and Notebook were used to provide a data exploration mechanism, model generation and experimentation. Here, we also used deep learning algorithms such as TensorFlow and Keras. All these experiments were performed on Windows 11, which supports all these software applications conveniently.

All the experiments, training, testing and validation were performed on the preprocessed images dataset (Uttaranchal University, Dehradun). To ensure optimal performance convergence and accuracy, the model was trained over 200 iterations and implemented different CNNs like ResNet50V2, ResNet152V2, InceptionV3, Xception, and MobileNetV2 to better understand the model and for better identification of different departments of the University.

5. Result and Discussion

This part covers the evaluation of the model's performance, including the tradeoff between accuracy and speed. It also examines the model's robustness to variations, limitations, and challenges. We used the GPU to conduct this experiment because using the CPU would take more than two hours for each iteration. Additionally, with each repetition, the accuracy increased by 5 to 7 percent. Figure 7 demonstrates the precision of the CNN model methods, including ResNet50V2, ResNet152V2, InceptionV3, Xception, and MobileNetV2. After undergoing forty iterations, we obtained a range of accuracy between 93 and 99 percent. Notably, the Xception CNN model demonstrated the highest level of accuracy, attaining an impressive 99.20 percent. To achieve a score of 99.20 and accurately forecast the content of photographs using the testing data, we train the Xception model for a maximum of 200 iterations.

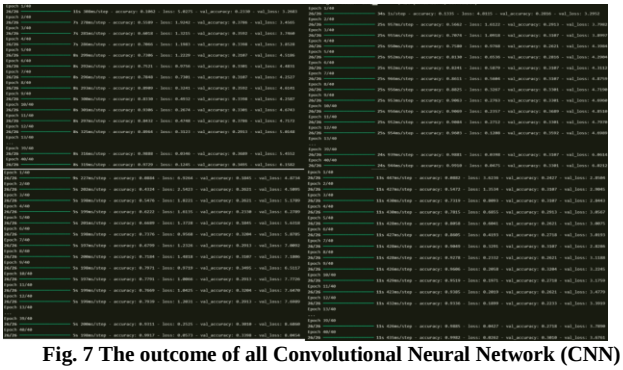

Fig. 7 The outcome of all Convolutional Neural Network (CNN) methods

5.1. Model Performance

The pictures in the Uttaranchal University dataset may be classified using several Convolutional Neural Network (CNN) designs. The accuracy rate of 99.82% attained by Xception is attributed to the use of depth-wise separable convolutions, which enable efficient extraction of features and learning of representations. The accuracy of InceptionV3 is 99.17%,

making it the algorithm with the best accuracy in identifying complex patterns and structures due to its inception modules. The accuracy of ResNet50V2 and ResNet152V2 is 98.88% and 99.10%, respectively, due to the efficacy of residual connections in training deeper networks and resolving the problem of vanishing gradients. Nevertheless, as the model grows more complex, the advantages in terms of performance become less apparent. As the model becomes more advanced, the visibility of efficiency savings diminishes. The accuracy of the MobileNetV2 model is 93.71%, which indicates the complexity and precision of the model. The accuracy of the model may be ascribed to its lightweight nature. A multitude of scholars are diligently striving to discover the most optimal design for Convolutional Neural Networks (CNNs) that align with their objectives and resources. Xception and InceptionV3 surpass more sophisticated techniques of visual analysis in terms of performance. The training and validation accuracy and loss of all CNN algorithms are shown in Figure 8. The Xception model obtains the best accuracy among all the CNN methods.

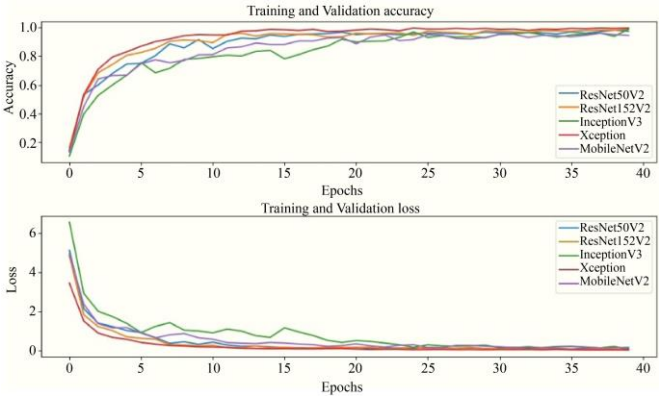

Fig. 8 Demonstrates the training and validation accuracy and loss for all CNN algorithms

5.2. Performance Analysis

The evaluation of the suggested method demonstrates convincing outcomes across several Convolutional neural network designs. The ResNet50V2 model demonstrated a remarkable accuracy of 98.88% in correctly categorizing photos from the Uttaranchal University dataset, highlighting its usefulness. ResNet152V2 had a little greater accuracy of 99.10%, highlighting the advantages of using deeper networks to capture intricate characteristics.

InceptionV3 surpassed both ResNet50V2 and ResNet152V2, attaining a remarkable accuracy of 99.17%. The use of inception modules in InceptionV3 improved the extraction of multi-scale features, enhancing the model's ability to capture complicated patterns and structures in the pictures more efficiently.

The Xception model had the greatest level of accuracy compared to all other models, with an exceptional accuracy rate of 99.82%. The significance of depth-wise separable convolutions in facilitating effective feature extraction and representation learning is emphasized by this. Figure 9 demonstrates the prediction results of the CNN model on the testing photos dataset, serving as an evaluation of its performance. A comparison of the accuracy of each image processing technique is shown in Table 1.

Table 1. The accuracy of image processing method

Model	Accuracy (in %)
ResNet50V2	98.88
ResNet52V2	99.10
InceptionV3	99.17
Xception	99.82
MobileNetV2	93.71

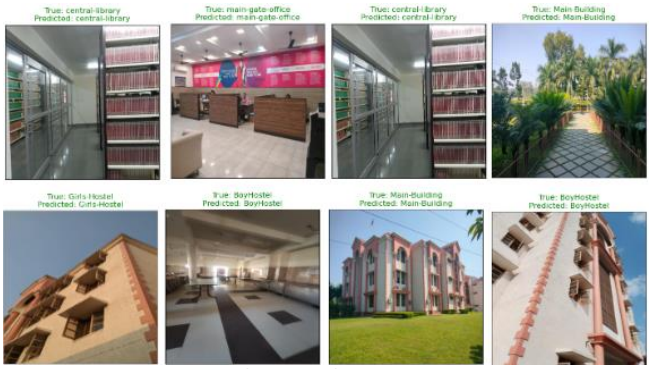

Fig. 9 The forecast of the testing photos.

Nevertheless, MobileNetV2 demonstrated a somewhat worse performance, with an accuracy rate of 93.71%. Although MobileNetV2 is specifically optimized for mobile and edge devices, its performance may have been compromised due to the balance between model complexity and accuracy.

6. Discussion

Most of the current research in image classification has opted for Convolutional Neural Networks (CNNs) as the neural network architecture to achieve the job. The feature layer gathers pertinent information via the process of convolution and pooling. Both steps, namely reducing the amount of information to be processed and extracting the final critical information known as the image map or signature, are performed using a fully connected neural network for image classification. This approach necessitates a variable number of parameters, ranging from thousands to millions, and the architecture adapts based on the information being processed. It signifies that the solution is intricate and might potentially affect the performance. Therefore, we suggest using this method with a maximum of 11000 parameters and a straightforward/static design, resulting in enhanced accuracy. CNN-based approaches can categorize images that include a backdrop because of their emphasis on this information. Regarding the university image dataset test, Sharma and Phonsa's technique [10] has a lower accuracy compared to

ours. Additionally, their approach has fewer classes since the dataset only consists of ten classes, and the images are not of high dimensions. The dissimilarity in signature between different types of images, such as dogs and cats, may be attributed to foveation, which consistently results in good performance. Therefore, we recommend that individuals choose our technique when the number of classes in the picture is small and there is little background, even if the image dimensions are significant. To determine whether the system should use convolution/pooling or PCNN/foveation as the feature extraction layer, a preprocessing module should be included in the processing chain.

7. Limitations and Challenges

Despite promising results, the testing procedure has significant drawbacks. Selecting hyperparameters for each CNN architecture proved difficult. Changes to hyperparameters like learning rate, batch size, and optimizer settings affected model convergence and performance. The dataset's size and variability made it challenging to depict several categories and their changes. Few data augmentation techniques were utilized to solve this challenge, but more complicated ones may increase the model's generalization and resilience. Computer resources limited the capacity to examine more complex designs or apply assembling procedures. Although the selected models performed well, ensemble techniques or larger ensemble models may improve classification accuracy. Another restriction is the likelihood of dataset bias from Uttaranchal University, Dehradun pictures' context. The dataset may not include all real-world photographs or biasing model predictions. The assessment measures solely considered classification accuracy, ignoring precision, recall, and F1-score. A comprehensive evaluation method may consider several factors that may improve future research by implementing more enhanced assessments that improve models' effectiveness. The recommended strategy produces promising results while addressing limits and challenges that improve real-world image categorization tasks and increase models' performance.

8. Conclusion

This study demonstrates the effectiveness of deep learning architectures in categorizing images from educational institutions like Uttaranchal University and Dehradun, which can aid in developing image classification systems for academic research and campus security. Researchers used cutting-edge designs like ResNet50V2, ResNet152V2, InceptionV3, Xception, and MobileNetV2, with Xception achieving the highest accuracy rate of 99.82%. The study identified CNN architectures' strengths, weaknesses, and shortcomings, allowing researchers to select models based on their objectives and limitations. The study also suggests that future research should explore aspects like ensemble learning, transfer learning, domain adaptability, and real-world deployment to improve the practicality of image categorization. The findings suggest that deep CNNs can effectively classify images from educational institutions, paving the way for robust image classification systems for campus security and academic research.

Declarations

Acknowledgement

All authors have read and agreed to the published version of the manuscript.

Funding

The authors are thankful for the support of Uttaranchal University, Doon University and Graphic Era (deemed to be) University.

References

[1] Kaiming He et al., "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, pp. 770-778, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[2] François Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," *2017 IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp. 1800-1807, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[3] Mark Sandler et al., "MobileNetV2: Inverted Residuals and Linear Bottlenecks," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 4510-4520, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[4] Christian Szegedy et al., "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, pp. 4278- 4284, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[5] Hamid Rezatofighi et al., "Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp. 658-666, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[6] Florian Debrauwer, EfficientNet | Rethinking Model Scaling for Convolutional Neural Networks, Medium, 2022. [Online]. Available: https://medium.com/to-cut-a-long-paper-short/efficientnet-rethinking-model-scaling-for-convolutional-neural-networks-eec0b2238b36

[7] Andrew G. Howard et al., "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," *Arxiv*, pp. 1-9, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[8] Wei Liu et al., "SSD: Single Shot MultiBox Detector," *Computer Vision – ECCV 2016: 14th European Conference*, Amsterdam, Netherlands, pp. 21-37, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[9] Karen Simonyan, and Andrew Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *Arxiv*, pp. 1-14, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[10] Ahmad Rofiqul Muslikh, De Rosal Ignatius Moses Setiadi, and Arnold Adimabua Ojugo, "Rice Disease Recognition Using Transfer Learning Xception Convolutional Neural Network," *Journal of Information Engineering*, vol. 6, no. 4, pp. 1535-1540, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[11] R. Vaitheeshwari, V. Sathiesh Kumar, and S. Anubha Pearline, "Design and Implementation of Human Safeguard Measure Using Separable Convolutional Neural Network Approach," *Computer Vision and Image Processing, Communications in Computer and Information Science*, vol. 1148, pp. 319-330, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[12] Tahmida Mahmud et al., "Prediction and Description of Near-Future Activities in Video," *Computer Vision and Image Understanding*, vol. 210, pp. 1-12, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[13] Huiren Tian et al., "An LSTM Neural Network for Improving Wheat Yield Estimates by Integrating Remote Sensing Data and Meteorological Data in the Guanzhong Plain, PR China," *Agricultural and Forest Meteorology*, vol. 310, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[14] Marco Klaiber, and Jonas Klopfer, "A Systematic Literature Review on SOTA Machine Learning-Supported Computer Vision Approaches to Image Enhancement," *Journal of Computer Science and Information*, vol. 15, no. 1, pp. 21-31, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[15] Ibrahim Goni, Asabe Sandra Ahmadu, and Yusuf Musa Malgwi, "Image Processing Techniques and Neuro-Computing Algorithms in Computer Vision," *Advances in Networks*, vol. 9, no. 2, pp. 33-38, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[16] Shuai Teng et al., "Structural Damage Detection Based on Convolutional Neural Networks and Population of Bridges," *Measurement*, vol. 202, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[17] Aya M. Shaaban, Nancy M. Salem, and Walid I. Al-Atabany, "A Semantic-Based Scene Segmentation Using Convolutional Neural Networks," *AEU - International Journal of Electronics and Communications*, vol. 125, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[18] Jianxun Lian et al., "Towards Better Representation Learning for Personalized News Recommendation: A Multi-Channel Deep Fusion Approach," *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pp. 3805-3811, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[19] Xingxun Jiang et al., "DFEW: A Large-Scale Database for Recognizing Dynamic Facial Expressions in the Wild," *Proceedings of the 28th ACM International Conference on Multimedia*, New York, United States, pp. 2881-2889, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[20] Gao Huang et al., "Densely Connected Convolutional Networks," *2017 IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp. 2261-2269, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[21] Mingxing Tan, and Quoc V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," *Arxiv*, pp. 1-11, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[22] Min Lin, Qiang Chen, and Shuicheng Yan, "Network in Network," *Arxiv*, pp. 1-10, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[23] Matthew D. Zeiler, and Rob Fergus, "Visualizing and Understanding Convolutional Networks," *Computer Vision – European Conference on Computer Vision 2014, Lecture Notes in Computer Science*, vol. 8689, pp. 818-833, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[24] Hugo Touvron et al., "Fixing the Train-Test Resolution Discrepancy," *Arxiv*, pp. 1-14, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[25] Dan Claudiu Cireşan et al., "Deep, Big, Simple Neural Nets for Handwritten Digit Recognition," *Neural Computation*, vol. 22, no. 12, pp. 3207-3220, 2010. [CrossRef] [Google Scholar] [Publisher Link]

[26] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the Knowledge in a Neural Network," *Arxiv*, pp. 1-9, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *Medical Image Computing and Computer-Assisted Intervention International Conference on Medical Image Computing and Computer-Assisted Intervention, Lecture Notes in Computer Science*, vol. 9351, pp. 234-241 2015. [CrossRef] [Google Scholar] [Publisher Link]

[28] N.L.W. Keijsers, "Neural Networks," *Encyclopedia of Movement Disorders*, pp. 257-259, 2010. [CrossRef] [Publisher Link]

[29] Yunchao Gong et al., "Multi-scale Orderless Pooling of Deep Convolutional Activation Features," *Computer Vision – European Conference on Computer Vision 2014, Lecture Notes in Computer Science*, vol. 8695, pp. 392-407, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[30] J.R.R. Uijlings et al., "Selective Search for Object Recognition," *International Journal of Computer Vision*, vol. 104, pp. 154-171, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[31] Andrej Karpathy et al., "Large-Scale Video Classification with Convolutional Neural Networks," *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, pp. 1725-1732, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[32] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," *Arxiv*, pp. 1-8, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[33] Arnab Ghosh et al., "Multi-Agent Diverse Generative Adversarial Networks," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 8513-8521, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[34] Ross Girshick, "Fast R-CNN," *2015 IEEE International Conference on Computer Vision*, Santiago, Chile, pp. 1440-1448, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[35] Hyeonseob Nam, and Bohyung Han, "Learning Multi-Domain Convolutional Neural Networks for Visual Tracking," *2016 IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, pp. 4293-4302, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[36] Yonis Gulzar, "Fruit Image Classification Model Based on MobileNetV2 with Deep Transfer Learning Technique," *Sustainability*, vol. 15, no. 3, pp. 1-14, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[37] Ke Dong et al., "MobileNetV2 Model for Image Classification," *2020 2nd International Conference on Information Technology and Computer Application*, Guangzhou, China, pp. 476-480, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[38] Slamet Riyadi, Febriyanti Azahra Abidin, and Nia Audita, "Comparison of ResNet50V2 and MobileNetV2 Models in Building Architectural Style Classification," *2024 International Conference on Intelligent Systems and Computer Vision*, Fez, Morocco, pp. 1-8, 2024. [CrossRef] [Google Scholar] [Publisher Link]