ADVANCED DEEP LEARNING METHODOLOGIES FOR MULTI-LABEL SATELLITE IMAGE CLASSIFICATION

A PROJECT REPORT

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE
OF

MASTER OF TECHNOLOGY
IN
INFORMATION TECHNOLOGY

Submitted by

Tamal Barman

23/ITY/15

Under the supervision of

Dr. Seba Susan

DEPARTMENT OF INFORMATION TECHNOLOGY
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi 110042

June, 2025

DEPARTMENT OF MECHANICAL ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I, TAMAL BARMAN, 2k23/ITY/15 of M.Tech (Information Technology),hereby declare that the project Dissertation titled "ADVANCED DEEP LEARNING METHODOLOGIES FOR MULTI-LABEL SATELLITE IMAGE CLASSIFICATION" which is submitted by me to the Department of Information Technology, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of degree of Master of Technology, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma Associateship, Fellowship or other similar title or recognition.

Place: Delhi Tamal Barman

June 2025

DEPARTMENT OF MECHANICAL ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CERTIFICATE

I hereby certify that the Project Dissertation titled "ADVANCED DEEP LEARNING METHOD- OLOGIES FOR MULTI-LABEL SATELLITE IMAGE CLASSIFICATION" which is submitted by Tamal Barman, Roll No. – 23/ITY/15, Information Technology ,Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Bachelor of Technology, is a record of the project work carried out by the students under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi — Seba Susan

June 2025 — **SUPERVISOR**

DEPARTMENT OF MECHANICAL ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

ACKNOWLEDGEMENT

I wish to express my sincerest gratitude to Dr Seba Susan for her continuous guidance and mentorship that she provided me during the project. She showed me the path to achieve our targets by explaining all the tasks to be done and explained to me the importance of this project as well as its industrial relevance. She was always ready to help me and clear our doubts regarding any hurdles in this project. Without her constant support and motivation, this project would not have been successful.

Place: Delhi Tamal Barman

Date: June 2025

Abstract

Multi-label satellite image classification presents significant challenges in remote sensing applications, as aerial scenes frequently contain multiple concurrent elements such as "partly cloudy," "agriculture," and "roads." The complexity increases due to ambiguous training data that often leads to overfitted models in deep learning approaches. Addressing these challenges, we propose a comprehensive framework that integrates both convolutional neural networks (CNNs) and transformer-based architectures to achieve optimal classification accuracy while enabling efficient deployment on resource-constrained devices. Our methodology implements a dual-approach strategy, thoroughly evaluating both paradigms on multi-label remote sensing datasets.

The first component of our framework utilizes the lightweight MobileNetV2 architecture pre-trained on millions of ImageNet images, implementing transfer learning techniques for multi-label classification. We incorporate an effective preprocessing pipeline featuring haze removal algorithms to enhance image quality prior to classification. During training, we employ one-hot encoding for the multiple class labels associated with each satellite image, while dynamically adjusting the threshold for posterior class probabilities at the network output to optimize prediction accuracy. This approach balances computational efficiency with classification performance, making it suitable for deployment in environments with limited resources.

Concurrently, we investigate Vision Transformers (ViTs) as an alternative paradigm, leveraging their unique ability to capture long-range dependencies across image regions. Unlike CNNs that extract features through convolutional layers, ViTs divide images into patches and process them as token sequences, similar to language processing techniques. This fundamental architectural difference enables ViTs to capture broader contextual information and more detailed features across the entire image-a critical advantage when dealing with multi-label satellite imagery containing diverse categories, sizes, and spatial arrangements. We comprehensively evaluate six lightweight ViT variants: ViT-Small,

ViT-Tiny, ViT-Base, Swin-Tiny, DeiT-Tiny, and DeiT-Base, optimizing each model for the multi-label classification task.

Our findings demonstrate that carefully optimized lightweight models can achieve performance comparable to or exceeding more complex architectures while requiring substantially fewer computational resources. This has important implications for real-time satellite image analysis, environmental monitoring, agricultural assessment, and disaster response applications where deployment on edge devices with limited processing capabilities is necessary. The complementary strengths of CNN and transformer-based approaches suggest that hybrid architectures combining aspects of both paradigms may represent a promising direction for future research in multi-label satellite image classification.

Contents

List of Tables

List of Figures

Chapter 1

INTRODUCTION

1.1 Environmental Impact of Deforestation

Deforestation and land-use changes have emerged as critical environmental challenges globally in recent decades. These human-induced activities have resulted in substantial greenhouse gas emissions and significant alterations to regional climate patterns. Research indicates that forests currently function as vital carbon sinks, absorbing an estimated 16 billion metric tonnes of carbon dioxide annually and storing approximately 861 gigatonnes of carbon in their branches, leaves, roots, and soils [21]. When forests are cut down or damaged, they lose their ability to absorb carbon dioxide from the air and instead start releasing it, turning from helpful carbon absorbers into major sources of pollution. Approximately 4.8 billion tons of carbon dioxide are released into the atmosphere annually as a direct result of deforestation activities, contributing to 11–20% of global greenhouse gas emissions [9].

Primary drivers of these transformations include agricultural expansion (particularly for high-value cash crops like soybean and palm oil), livestock production (especially cattle ranching), infrastructure development, mining operations, and urbanization. Mining activities are particularly detrimental, causing direct habitat destruction at extraction sites while simultaneously contributing to broader environmental degradation through pollution and landscape modification. These combined activities contribute to extensive ecological degradation, with research suggesting that land-use changes have already caused ecological communities to lose an average of 13.6% of species globally [21].

Beyond carbon emissions, deforestation significantly disrupts hydrological cycles. Forests regulate atmospheric moisture through transpiration, with trees absorbing groundwater and releasing it into the atmosphere through their leaves. This process generates localized humidity and influences precipitation patterns across regions. When forests are removed, less water evaporates into the atmosphere, causing land to become drier and less stable, often leading to increased soil erosion, flooding, and in extreme cases, desertification. The Amazon rainforest exemplifies this vulnerability, having already lost nearly 17% of its original forest cover and approaching an ecological tipping point where the remaining forest may be insufficient to maintain regional hydrological cycles [9].

1.2 Significance of Remote Sensing in Environmental Monitoring

Remote sensing has become a commendable tool in environmental monitoring due to its ability to provide continuous, large-scale, and multi-temporal observations of the Earth's surface. Unlike traditional field-based surveys, which are limited by accessibility, labor intensity, and high operational costs, remote sensing technologies allow researchers to observe deforestation, land-use change, urban expansion, and other ecological processes over vast and often inaccessible areas with high spatial and temporal resolution [20]. Satellite sensors, in particular, offer a consistent and repetitive data acquisition platform, facilitating long-term environmental monitoring and change detection at both local and global scales [29]. Understanding how forests change over time is crucial, especially in the

Figure 1.1: Satellite imagery used to monitor deforestation and land-use change.

face of deforestation and shifting land-use patterns. Remote sensing has become a key tool in monitoring these changes, allowing researchers to map forest cover, detect early signs of degradation, and measure the pace and scale of forest loss. Satellites like Landsat [29], Sentinel [7], and MODIS [17] gather spectral, spatial, and temporal data that help differentiate between primary forests, plantations, and degraded areas.

The ability of these systems to capture multispectral and hyperspectral imagery plays a big role in identifying different vegetation types, canopy densities, and overall forest health. When combined with GIS (Geographic Information Systems), this data becomes even more valuable. It allows scientists and decision-makers to overlay satellite observations with other important layers—like elevation, soil characteristics, or demographic data—which improves the depth of environmental assessments and strengthens land-use planning strategies.

One especially impactful use of remote sensing is in spotting illegal deforestation early. High-resolution imagery and change detection algorithms can reveal unapproved land clearing or logging activity in near real-time. This gives authorities and conservation groups the ability to act quickly and enforce protection laws. Remote sensing also supports global efforts such as the REDD+ initiative (REDD+, 2021), helping countries account for carbon emissions tied to forest loss and receive funding for conservation through

performance-based incentives.

Of course, remote sensing isn't without its challenges. Optical imagery can be blocked by clouds, image interpretation can be tricky in complex terrains, and ground-truth data is often needed to validate findings. Still, continuous advancements in sensor technology, open-access platforms, and smarter data processing (Lu et al., 2007) are making these tools more accurate and accessible than ever[20].

Figure 1.2: Multi-label images for monitoring deforestation and land-use changes .

In remote sensing, the classification of satellite imagery is vital for extracting meaningful information about land-use and land-cover (LULC) features. Traditional classification techniques in remote sensing typically operate under the assumption that each pixel belongs exclusively to a single class. While this simplification may be computationally efficient, it often falls short in representing the complexity of real-world environments. In many scenarios—especially in heterogeneous regions such as tropical forests, agricultural mosaics, or peri-urban zones—multiple land-cover types may coexist within the same image segment. As a result, multi-label classification becomes essential for accurately capturing and interpreting such diverse spatial compositions.

Unlike single-label classification, the multi-label approach permits assigning more than one class label to a given image or pixel, thereby offering a more comprehensive view of the landscape. For example, a satellite image of a tropical area might simultaneously show patches of dense forest, cleared land, and cultivated fields. In these cases, multi-label classification helps ensure that each feature is correctly identified, reducing the likelihood of misclassification and enhancing the usefulness of the data for downstream applications.

The significance of adopting this method extends beyond accuracy. As shown in Fig. 1.2, multi-label classification provides richer insights into land-cover transitions, which are valuable for environmental monitoring, policy-making, and disaster risk assessment. Identifying areas with overlapping features—such as degraded forests being overtaken by agriculture—can guide more targeted conservation efforts and inform land-use planning decisions.

Moreover, this classification strategy enhances the performance of machine learning models by aligning with the inherently complex nature of ecological data. Methods like binary relevance, classifier chains, and neural network-based adaptations allow the model to learn label dependencies and correlations more effectively, thereby improving robustness and generalization.

In summary, multi-label classification offers a more realistic and powerful framework for analyzing remote sensing data, particularly in dynamic and multi-faceted landscapes. Its use not only advances environmental understanding but also supports more informed decisions in the domains of sustainability, land management, and resource planning.

1.3 Advancements in Deep Learning for Image Classification

Over the past decade, deep learning has dramatically reshaped the landscape of image classification by enabling models to automatically learn and extract hierarchical features directly from raw image data. One of the most influential architectures driving this transformation has been the Convolutional Neural Network (CNN), which has consistently demonstrated strong performance across a wide spectrum of computer vision tasks. From basic object recognition and image segmentation to more complex applications like remote sensing image analysis, CNNs have proven to be highly effective tools [20, 17].

The strength of CNNs lies in their ability to capture local spatial features using convolutional layers that apply filters across the input image. These layers are often followed by pooling operations, which help in reducing dimensionality while preserving key structural information. As the network goes deeper, it builds increasingly abstract representations of the input, allowing it to recognize intricate patterns and textures.

Despite their widespread success, CNNs do have intrinsic limitations. One of the key challenges is their restricted ability to capture long-range dependencies and global contextual information. The receptive field of a CNN grows with depth, but this growth is incremental. As a result, CNNs may fail to effectively model relationships between distant regions in an image—an important capability when working with large, complex scenes like those found in satellite imagery [22]. Additionally, the use of fixed-size kernels may cause the network to overlook fine details or global spatial structures that fall outside the local receptive area.

To overcome these constraints, a new class of models known as Vision Transformers (ViTs) has emerged, bringing a paradigm shift in how visual data is processed. Drawing inspiration from the groundbreaking success of transformer architectures in natural language processing, ViTs adapt the attention mechanism to image analysis. They work by first dividing an image into a sequence of fixed-size patches. Each patch is then flattened and linearly embedded into a vector, effectively treating image patches as tokens, akin to words in a sentence. These tokens are passed through multiple transformer layers, where self-attention mechanisms enable the model to capture relationships and dependencies across the entire image context [5].

What sets ViTs apart is their inherent ability to model global interactions right from the initial layers, something CNNs typically struggle with. This attribute allows ViTs to understand the broader context of an image more effectively, making them well-suited for analyzing satellite images where objects and features are often distributed across large spatial extents. Furthermore, ViTs are highly scalable and can adapt to varying data sizes and structures, which is crucial in remote sensing tasks that often involve massive and diverse datasets.

Recent studies have shown that, when provided with adequate training data, Vision Transformers can outperform traditional CNNs on several image classification benchmarks. Their flexibility, robustness, and improved capability in capturing global image semantics make them a compelling choice for next-generation image analysis in environmental monitoring, land use classification, and other remote sensing applications [27, 28].

Chapter 2

LITERATURE REVIEW

As forests continue to vanish across the globe at an alarming pace, understanding the broader consequences of this loss has become increasingly urgent. Deforestation does not just remove trees—it disrupts ecosystems, accelerates climate change, and threatens countless plant and animal species. If land-use patterns and energy consumption practices remain unchanged, the combined pressures of climate change and large-scale forest loss could result in widespread biodiversity decline, destabilizing the delicate balance of nature and directly impacting human well-being.

In response to these growing environmental concerns, Remote Sensing (RS) technologies have emerged as essential tools in the global effort to monitor and address deforestation. By capturing and analyzing satellite imagery over time, researchers and policymakers can gain valuable insights into how forest cover is changing across different regions. This visual and data-driven perspective allows for more precise tracking of forest degradation, land-use shifts, and environmental stressors [1, 10, 11].

What makes remote sensing particularly powerful is its ability to provide consistent, up-to-date, and wide-scale information. This enables governments, conservationists, and organizations to not only assess the current state of forests but also to forecast future risks and prioritize areas most in need of protection. In essence, continuous monitoring through RS supports the development of smarter, more targeted conservation strategies—ones that can effectively reduce the impact of deforestation and promote sustainable land management practices.

The rise of satellite technology in the 1970s, classifying remote sensing imagery has been a key area of research. While the tools have evolved over time, the basic approach has remained largely the same: gather satellite data, extract important features, apply classification techniques, and produce thematic maps that show different types of land cover. [8]. Traditional feature engineering approaches relied predominantly on fundamental image processing techniques, including image filtering, clustering algorithms, Principal Component Analysis (PCA), and feature selection methods. However, these conventional models frequently demonstrated inadequate precision and encountered difficulties in capturing the complex interrelationships inherent in satellite imagery data .

The early 2000s witnessed a paradigmatic shift as remote sensing methodologies began to integrate with emerging fields such as deep learning and computer vision. This interdisciplinary convergence catalyzed the development of sophisticated machine learning techniques specifically tailored for remote sensing image analysis [2]. Concurrent advancements in computational capabilities have rendered deep convolutional neural networks (CNNs) viable for large-scale processing and interpretation of remote sensing imagery. These deep learning architectures demonstrate exceptional capacity for comprehending intricate spatial relationships within satellite images, thereby significantly enhancing clas-

sification accuracy relative to traditional methodological approaches [31].

Despite significant advances in deep learning, classifying satellite images remains a challenging task—especially when it comes to multi-label classification, where images often contain multiple overlapping features or land cover types. Numerous satellite images simultaneously encompass multiple land cover types and atmospheric conditions—for instance, regions may be characterized as "partly cloudy" while concurrently exhibiting "agricultural" features and "transportation infrastructure." The accurate identification of all applicable scene labels in such complex contexts remains challenging, especially when certain land cover categories are underrepresented in training datasets [16].

The research community has implemented various pre-trained CNN architectures for satellite image classification. These implementations typically employ transfer learning methodologies, wherein pre-trained models are fine-tuned for the specific task of satellite image classification. This approach substantially reduces computational requirements and training duration compared to developing models from initialization. Among CNN architectures, MobileNetV2 warrants particular consideration for resource-constrained applications due to its efficient design (comprising 28 layers and approximately 0.2 million parameters), which incorporates depthwise separable convolutions and residual bottleneck layers [3].

Concurrent with CNN advancements, Vision Transformer (ViT) models have emerged as viable alternatives to conventional CNN-based approaches for processing complex image datasets. ViTs process images as sequences of patches, employing self-attention mechanisms to capture long-range dependencies and globally relevant information [5, 19]. This architectural paradigm facilitates more nuanced interpretation and segmentation of complex image features. Chen et al. demonstrated the efficacy of ViT-based models in segmenting images and capturing fine details across entire scenes , while Yao et al. extended ViT applications to effective land-cover change mapping utilizing high-resolution imagery. Furthermore, Kaselimi et al. employed ViT variants for precise deforestation mapping, surpassing conventional CNN models through the application of self-attention mechanisms that prioritize the most relevant visual information [18, 30].

Recent developments in artificial intelligence have fostered the creation of lightweight ViT architectures that maintain effectiveness while enhancing computational efficiency. These streamlined models incorporate fewer parameters than baseline ViT architectures and are suitable for deployment in resource-constrained computational environments, including mobile platforms. The Data-Efficient Transformer (DeiT) exemplifies such lightweight ViT models and has been successfully implemented for the classification of horticultural plantations using satellite imagery [25].

In this study, we explore and compare the capabilities of Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) in the context of multi-label classification of satellite imagery. The analysis focuses on a diverse set of deep learning models, including MobileNetV2, ResNet-50, Inception-v3, DenseNet-121, and the standard Vision Transformer (ViT). Alongside these, we conduct a detailed investigation of six lightweight ViT variants: ViT-Small, ViT-Tiny, ViT-Base, Swin-Tiny, DeiT-Tiny, and DeiT-Base.

Our experimental framework applies these models to satellite images of the Amazon rainforest, a region characterized by complex land cover and rich ecological diversity. Each image may correspond to multiple land-use categories, with the dataset containing annotations across seventeen distinct labels. This multi-label nature presents a challenging classification scenario, requiring models to identify and assign multiple relevant tags to each image.

To measure the performance of each model architecture, we employ the F-beta score, a widely used evaluation metric that balances precision and recall. This enables a comprehensive assessment of how effectively each model captures the diverse and overlapping features present in the satellite imagery, providing insight into the strengths and limitations of different deep learning approaches in remote sensing applications.

Chapter 3

DEEP LEARNING ARCHITECTURES

3.1 ResNet-50

ResNet-50 is a widely recognized convolutional neural network (CNN) architecture that has significantly influenced the field of deep learning and computer vision. Developed by Kaiming at Microsoft Research Asia, ResNet-50 was introduced in 2015 as part of the Residual Networks (ResNets) family [13]. The architecture consists of 50 layers and is designed around the concept of residual learning, which enables the training of much deeper neural networks by addressing the vanishing gradient problem. By incorporating identity shortcut connections, ResNet-50 allows gradients to flow more effectively through the network during backpropagation, facilitating improved convergence and overall performance. This innovation has made ResNet-50 a foundational model for a wide range of visual recognition tasks, including object detection, image classification, and remote sensing applications. [13]. ResNet-50 addressed a critical issue in training deep neural networks: as the network depth increases, performance degrades, not due to overfitting but because of optimization difficulties such as the vanishing gradient problem. The authors proposed that deeper models should theoretically perform at least as well as shallower ones, provided that the added layers could learn identity mappings.

3.1.1 Architectural Innovation: Residual Learning

The core innovation of ResNet-50 lies in its use of residual blocks and skip connections. As illustrated in Fig. 3.1, residual learning reformulates the layers as learning residual functions with reference to the layer inputs, rather than directly trying to learn unreferenced functions.

Mathematically, instead of a desired underlying mapping denoted as $H(x)$, the residual block aims to learn the residual function $F(x) = H(x) - x$, which leads to the reformulation:

$$H(x) = F(x) + x \tag{3.1}$$

Here, x is the input, $F(x)$ is the residual mapping learned by the block (a stack of convolutional layers), and $H(x)$ is the output. This formulation allows gradients to be backpropagated more directly, alleviating the vanishing gradient issue and enabling successful training of very deep networks.

3.1.2 Detailed Architecture of ResNet-50

The ResNet-50 architecture begins with a 7×7 convolutional layer with stride 2, followed by batch normalization, ReLU activation, and a 3×3 max pooling layer. These operations are designed to capture low-level spatial features in the input image.

Following this, the network is composed of four stages, each containing multiple residual bottleneck blocks. As shown in Fig. 3.2, each bottleneck block implements three convolutional layers:

1. A 1×1 convolution for reducing dimensionality
2. A 3×3 convolution for processing features
3. A 1×1 convolution for restoring dimensionality

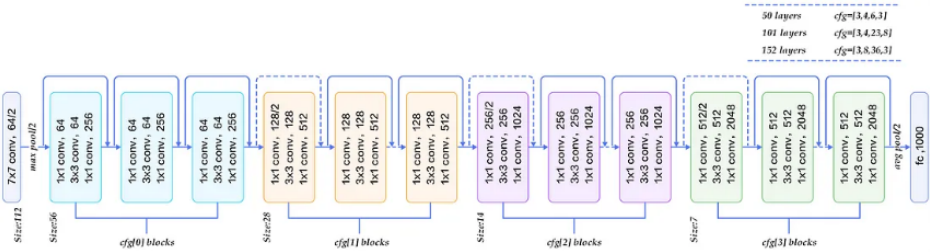

Figure 3.1: ResNet-50 architecture overview adapted from [13].

Each of these is followed by batch normalization and ReLU activation (except the final convolution). The identity (skip) connection adds the input x directly to the output of the block:

$$\text{Output} = F(x, \{W_i\}) + x \tag{3.2}$$

where $\{W_i\}$ represents the weights of the layers in the residual function $F(x)$.

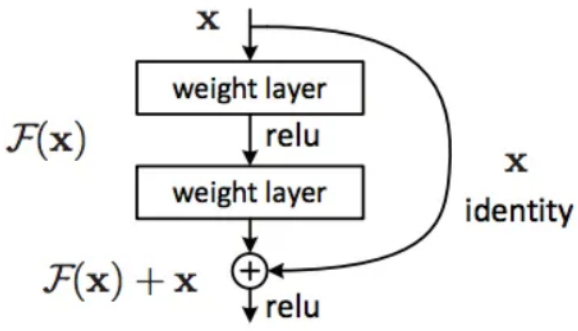

Figure 3.2: Skip connections enable identity mapping and improve gradient flow.

To adjust dimensions when needed (such as during downsampling or channel mismatch), a projection shortcut using a 1×1 convolution is applied to x, replacing the identity shortcut:

$$\text{Output} = F(x, \{W_i\}) + W_s x \tag{3.3}$$

where W_s denotes the projection weights.

After the four stages of residual blocks, ResNet-50 ends with a global average pooling layer that reduces each feature map to a single value, followed by a fully connected layer with softmax activation for classification tasks.

3.2 MobileNetV2

MobileNetV2 is a lightweight convolutional neural network architecture tailored for applications where computational resources are limited, such as mobile devices and embedded systems. It builds on the original MobileNet design by introducing two key innovations—*inverted residuals* and *linear bottlenecks*—which together enable a drastic reduction in parameters and multiply–accumulate operations without sacrificing accuracy [23].

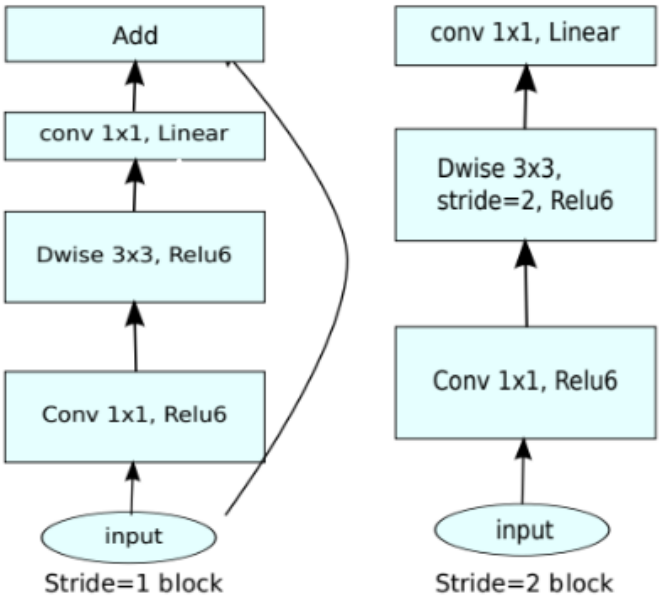

Figure 3.3: MobileNetV2 architecture adapted from [23] .

The core innovation in MobileNetV2 lies in how each block is structured to optimize both efficiency and representational capacity. Each block consists of three major stages—expansion, depthwise convolution, and projection—which can be mathematically represented to better understand how features are processed.

In the **expansion phase**, a 1×1 convolution is applied to expand the dimensionality of the input feature map $\mathbf{X} \in \mathbb{R}^{H \times W \times C_{in}}$ into a higher-dimensional space:

$$\mathbf{X}_{\text{exp}} = \sigma(\mathbf{W}_e * \mathbf{X}) \tag{3.4}$$

where $\mathbf{W}_e$ is the weight tensor for the expansion layer and σ is the non-linear activation function (ReLU6 is commonly used).

The **depthwise convolution** then performs a separate convolution for each channel, significantly reducing computation:

$$\mathbf{X}_{\text{dw}} = \sigma(\mathbf{W}_{\text{dw}} \odot \mathbf{X}_{\text{exp}}) \tag{3.5}$$

Here, $\odot$ denotes the channel-wise (depthwise) convolution, and $\mathbf{W}_{\text{dw}}$ contains the depthwise filters.

Finally, in the **projection phase**, a 1×1 linear convolution projects the output back to a lower-dimensional space without using an activation function:

$$\mathbf{X}_{\text{proj}} = \mathbf{W}_p * \mathbf{X}_{\text{dw}} \tag{3.6}$$

where $\mathbf{W}_p$ is the projection weight tensor. The use of a linear activation in this step helps preserve information that might otherwise be lost in a narrow bottleneck.

MobileNetV2 also uses residual connections when the input and output dimensions match:

$$\mathbf{Y} = \mathbf{X} + \mathbf{X}_{\text{proj}} \tag{3.7}$$

The introduction of residual connections in deep networks plays a crucial role in improving gradient flow during backpropagation. In the case of ResNet-50, this architectural innovation enhances the network's ability to learn deep feature hierarchies without encountering vanishing gradient issues. By allowing information to bypass certain layers and flow directly to deeper parts of the network, these connections strengthen the model's representational capacity while adding minimal computational overhead.

Similarly, MobileNetV2 incorporates a range of design choices aimed at achieving both computational efficiency and strong performance. Its architecture includes inverted residual blocks with linear bottlenecks, which help reduce the number of parameters and improve memory usage. These structural optimizations allow MobileNetV2 to deliver accurate results even in constrained environments. As a result, it is particularly well-suited for real-time image classification tasks and deployment on resource-limited devices, such as mobile phones and embedded systems.

3.3 DenseNet-121

DenseNet-121 is a deep convolutional neural network known for its innovative connectivity design. Presented by Gao Huang and colleagues in their influential paper *"Densely Connected Convolutional Networks"* [15], this architecture breaks away from traditional network structures by establishing direct connections between each layer and all subsequent layers. This dense pattern of connectivity helps to enhance the flow of information and gradients throughout the network, promoting more efficient feature reuse and reducing redundancy. As a result, DenseNet-121 achieves impressive parameter efficiency while maintaining strong performance, which has made it a popular choice for a wide range of computer vision applications.

3.3.1 Architectural Design

Unlike traditional convolutional networks that arrange layers in a simple sequential order, DenseNet-121 uses a unique feed-forward design. In this structure, each layer receives input not only from the immediate preceding layer but also from the outputs of all earlier layers within the same dense block. This approach encourages continuous reuse of features throughout the network, leading to greater feature efficiency and minimizing redundancy.

The network is composed of four dense blocks, each containing multiple convolutional layers as shown in Fig 3.4. These blocks are separated by transition layers that consist of batch normalization, a 1×1 convolution for channel compression, and a 2×2 average

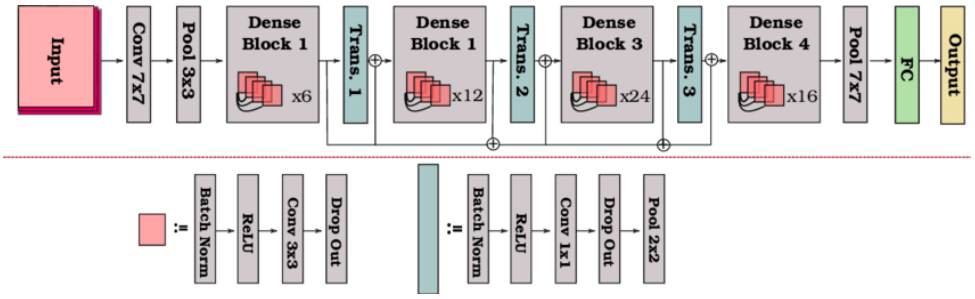

Figure 3.4: DenseNet-121 architecture adapted from[14] .

pooling operation for spatial downsampling. The number "121" denotes the total number of layers in the network, including convolutional, batch normalization, and fully connected layers.

A distinctive characteristic of DenseNet is its use of a fixed growth rate, which defines how many new feature maps each layer contributes. In DenseNet-121, the growth rate is typically set to 32, meaning each layer within a dense block outputs 32 new feature maps that are then concatenated with all previous feature maps.

Mathematically, the input to the ℓ-th layer in a dense block can be expressed as:

$$\mathbf{x}_\ell = H_\ell([\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{\ell-1}]) \tag{3.8}$$

where $[\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{\ell-1}]$ represents the concatenation of the feature maps produced by layers 0 to $\ell - 1$, and $H_\ell(\cdot)$ denotes the composite non-linear transformation (typically Batch Normalization, ReLU, and a 3×3 convolution) applied at layer ℓ.

The output dimensionality after the ℓ-th layer grows linearly with the number of layers in the block, governed by the growth rate k:

$$C_\ell = C_0 + k \cdot \ell \tag{3.9}$$

where C_0 is the number of input channels to the block, and C_ℓ is the number of output channels at layer ℓ.

Transition layers compress the feature maps between dense blocks using a 1×1 convolution followed by average pooling. The compression factor θ controls the reduction in channels:

$$C_{\text{out}} = \theta \cdot C_{\text{in}} \tag{3.10}$$

where $0 < \theta \leq 1$ (typically $\theta = 0.5$).

This dense connectivity pattern significantly improves parameter efficiency, encourages feature reuse, and enables the training of very deep models without overfitting or vanishing gradients. DenseNet-121 has thus become a popular choice for various classification and segmentation tasks.

3.4 Inception v3

Inception v3 is a deep convolutional neural network architecture that improves upon earlier Inception models by incorporating several enhancements aimed at increasing computational efficiency while preserving or even improving accuracy [24]. It was developed by Christian Szegedy et al. and has been widely adopted for image classification tasks due to its strong performance on the ImageNet dataset.

3.4.1 Architectural Design

The Inception architecture is characterized by its use of *Inception modules*, which apply multiple types of convolutions in parallel to the same input and concatenate their outputs. This design allows the network to capture features at multiple spatial scales simultaneously. Inception v3 enhances this idea by incorporating several advanced techniques:

- **Factorized Convolutions:** Large convolutional filters such as 5×5 are replaced with consecutive smaller filters (e.g., two 3×3 convolutions) to reduce computation.
- **Asymmetric Convolutions:** Filters are decomposed into $n \times 1$ followed by $1 \times n$ convolutions to further reduce computational complexity.
- **Auxiliary Classifiers:** Additional classifiers are added during training to improve gradient flow and act as regularizers.
- **Label Smoothing:** A regularization technique that prevents the model from becoming overconfident by softening the target labels.

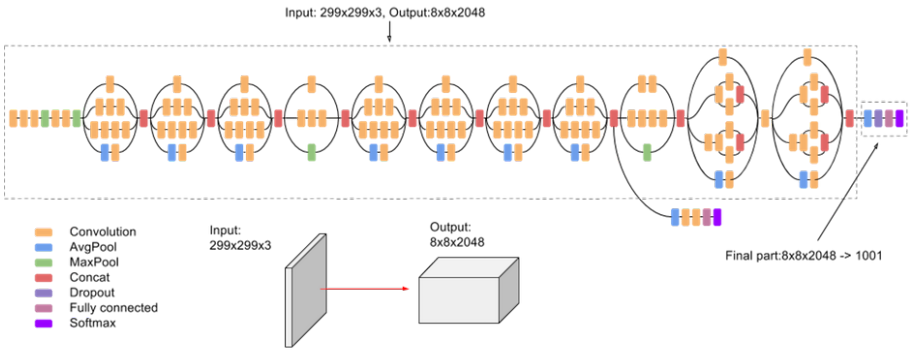

Figure 3.5: Inception v3 architecture adapted from[24] .

3.4.2 Mathematical Components

To reduce computational cost, a 5×5 convolution is factorized into two consecutive 3×3 convolutions. Assuming the number of input and output channels is the same, the cost reduction can be approximated as:

$$\text{Cost}_{5\times5} = 25C^2, \quad \text{Cost}_{3\times3\times2} = 18C^2 \tag{3.11}$$

This leads to a reduction in computational cost by approximately 28%, while retaining the same receptive field.

Inception v3 also applies asymmetric convolutions to further reduce complexity. A $n \times n$ convolution is approximated using:

$$f(x) = \text{Conv}_{1\times n}(\text{Conv}_{n\times1}(x)) \tag{3.12}$$

This technique not only reduces the number of operations but also introduces additional non-linearity, improving representational power.

Label smoothing is applied to the softmax output to regularize training by modifying the ground-truth distribution $q(k)$ as:

$$q'(k) = (1 - \epsilon) \cdot q(k) + \frac{\epsilon}{K} \tag{3.13}$$

where K is the number of classes and ϵ is a small constant (e.g., $\epsilon = 0.1$). This prevents the model from becoming overconfident and improves generalization.

3.5 Vision Transformer (ViT)

The Vision Transformer (ViT) introduces a fundamentally different approach to image recognition by utilizing the Transformer architecture, which was originally designed for natural language processing tasks. Unlike traditional convolutional neural networks (CNNs) that rely on local receptive fields and hierarchical feature extraction, ViT models the image as a sequence of patches and applies self-attention mechanisms to learn global context from the outset. This paradigm shift allows ViT to effectively capture long-range dependencies and complex spatial relationships within images, especially when trained on large-scale datasets [5].

3.5.1 Fundamental Design Principles

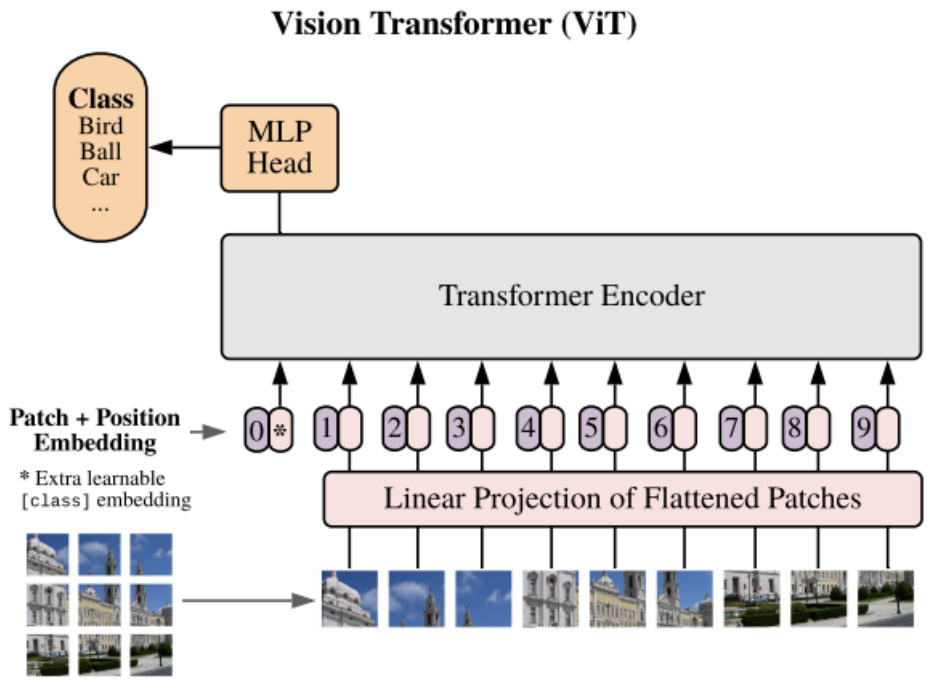

Figure 3.6: Architecture of Vision Transformer (ViT), showing patch division, embedding, Transformer layers, and final classification adapted from [6] .

At the heart of ViT is the idea of treating image patches similarly to words in a sentence. A given input image $x \in \mathbb{R}^{H \times W \times C}$ is divided into N fixed-size non-overlapping

patches of dimension $P \times P$, where $N = \frac{H \cdot W}{P^2}$. Each patch is then flattened and passed through a linear projection layer that transforms it into a D-dimensional embedding:

$$z_0^i = x_p^i E, \quad \text{for } i = 1, 2, \ldots, N \tag{3.14}$$

To incorporate positional information, which is not inherently captured by the attention mechanism, learnable positional encodings E_{pos} are added to the projected patch embeddings:

$$z_0 = [x_{\text{cls}}; z_0^1; z_0^2; \ldots; z_0^N] + E_{\text{pos}} \tag{3.15}$$

Here, x_{cls} is a special classification token appended at the beginning of the sequence. This token is designed to aggregate information from all other patches during training.

3.5.2 Transformer Encoder Structure

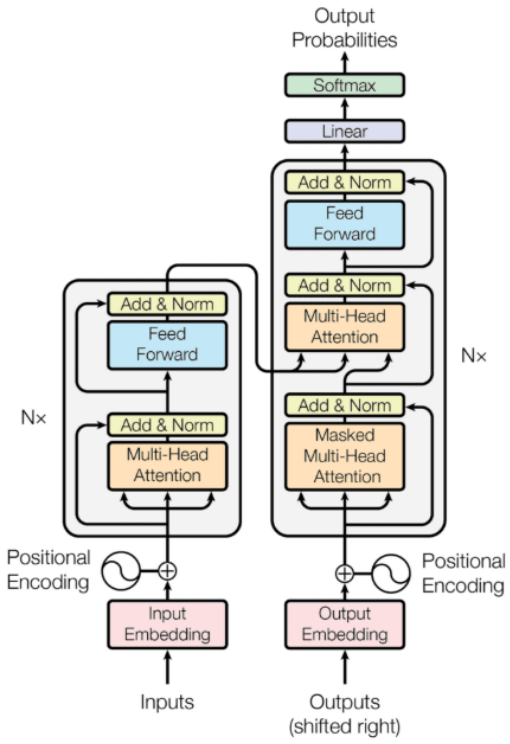

Figure 3.7: Transformer Encoder and Decoder adapted from [6]

The combined sequence of embeddings is passed through a stack of L identical Transformer encoder layers. Each layer comprises two main components: multi-head self-attention (MSA) and a position-wise feed-forward network (FFN). These layers are wrapped with residual connections and layer normalization (LN), ensuring training stability and effective gradient flow:

$$z'_l = \text{MSA}(\text{LN}(z_{l-1})) + z_{l-1} \tag{3.16}$$

$$z_l = \text{MLP}(\text{LN}(z'_l)) + z'_l \tag{3.17}$$

This architecture allows ViT to compute attention scores across all patches at once, capturing interactions between distant parts of the image.

3.5.3 Classification Head

Once the input passes through all Transformer layers, the output corresponding to the classification token is extracted and used for the final prediction. A simple fully connected layer followed by a softmax activation computes the class probabilities:

$$\hat{y} = \text{softmax}(W z_L^{[\text{CLS}]} + b) \tag{3.18}$$

where W and b are trainable weights and bias terms, and $z_L^{[\text{CLS}]}$ denotes the final embedding of the classification token after L layers.

3.6 Vision Transformer Variants

Since the introduction of the original Vision Transformer (ViT) architecture, several variants have emerged to cater to different trade-offs between accuracy, speed, and computational complexity. These variants primarily differ in model size, the number of parameters, and internal architectural configurations such as patch size and embedding dimensions. This section outlines some commonly used ViT variants and their distinctive characteristics.

3.6.1 ViT_tiny_patch16_224

The **ViT-Tiny** variant is a compact model designed for efficient inference on low-resource hardware. It uses 16×16 patch sizes and processes input images of size 224×224. With a smaller embedding dimension (typically 192) and fewer transformer blocks (e.g., 12 layers), it offers fast computation at the expense of slightly reduced accuracy. It is ideal for applications where speed and efficiency are critical.

3.6.2 ViT_small_patch16_224

ViT-Small provides a balance between efficiency and performance. Like ViT-Tiny, it uses a patch size of 16×16, but increases the embedding dimension to 384 and typically uses 12 transformer layers. This enhancement enables better representation learning without a significant increase in computational cost, making it suitable for mid-sized applications.

3.6.3 ViT_base_patch16_224

The **ViT-Base** model serves as a standard benchmark for many Vision Transformer studies. It uses 16×16 patches and 224×224 resolution images, with an embedding dimension of 768 and 12 transformer encoder layers. This model achieves strong performance on large-scale datasets like ImageNet while maintaining manageable training costs. It serves as a foundation for more advanced or larger ViT configurations.

3.6.4 Swin_tiny_patch4_window7_224

Swin-Tiny, or Swin Transformer Tiny, introduces a hierarchical transformer structure where self-attention is computed within local windows, and windows shift between layers. This patch-based model starts with 4×4 non-overlapping patches and utilizes a window size of 7×7. Such a design reduces computational complexity while preserving spatial

locality. Swin-Tiny achieves better accuracy than standard ViT-Tiny with similar or lower computational cost.

3.6.5 DeiT_tiny_patch16_224

The **Data-efficient Image Transformer (DeiT-Tiny)** is a distilled version of ViT-Tiny designed for data-efficient training. Like its counterpart, it uses 16×16 patch size and 224×224 image input, but it incorporates knowledge distillation by using a distillation token alongside the class token. This helps the model learn more effectively with fewer training samples, improving performance without increasing the model size.

3.6.6 DeiT_base_patch16_224

DeiT-Base extends the benefits of distillation to the base ViT model. It retains the same configuration as ViT-Base (patch size 16×16, image size 224×224, embedding size 768, 12 layers) while incorporating the distillation strategy. The result is a model that matches or surpasses ViT-Base performance with improved training data efficiency, especially beneficial when large-scale datasets are not available.

Comparison Summary

Table 3.1: Comparison of ViT and Transformer Variants

Model	Embedding Dim	Layers	Patch Size	Image Size
ViT-Tiny-P16-224	192	12	16×16	224×224
ViT-Small-P16-224	384	12	16×16	224×224
ViT-Base-P16-224	768	12	16×16	224×224
Swin-Tiny-P4-W7-224	Varies	Varies	4×4	224×224
DeiT-Tiny-P16-224	192	12	16×16	224×224
DeiT-Base-P16-224	768	12	16×16	224×224

These variants demonstrate how design adjustments—such as patch granularity, embedding size, or architectural strategies like windowed attention—can significantly impact performance, resource usage, and training efficiency. Selecting the right variant depends on the specific constraints and goals of a given application.

Chapter 4

METHODOLOGY

In this study, we address the problem of multi-label scene classification in Amazon rainforest satellite imagery by combining advanced preprocessing techniques with a lightweight convolutional backbone. First, we describe the dataset and its preparation; next, we introduce our haze-removal step; finally, we detail the design of the classification pipeline built upon MobileNetV2.

4.1 Dataset Preparation

We utilize the "Planet: Understanding the Amazon from Space" dataset which comprises 40 479 RGB images of size 256×256, each annotated with one or more of 17 possible labels (e.g., *agriculture*, *water*, *roads*,*etc*) as shown in Fig 4.1. These images, captured by sun-

Figure 4.1: Random samples containing the tags from the dataset

synchronous orbit satellites, provide a high-resolution view of land use and environmental conditions in the Amazon basin. To ensure consistency across all inputs, we convert every image to PNG format, then resize it to 224×224 pixels. Following this, each channel is normalized by

$$x' = \frac{x - \mu}{\sigma} \tag{4.1}$$

where μ and σ are the per-channel mean and standard deviation computed over the training split. We partition the dataset into training and test subsets using an 80:20 split and employ five-fold cross-validation to assess model robustness.

4.1.1 Haze Removal

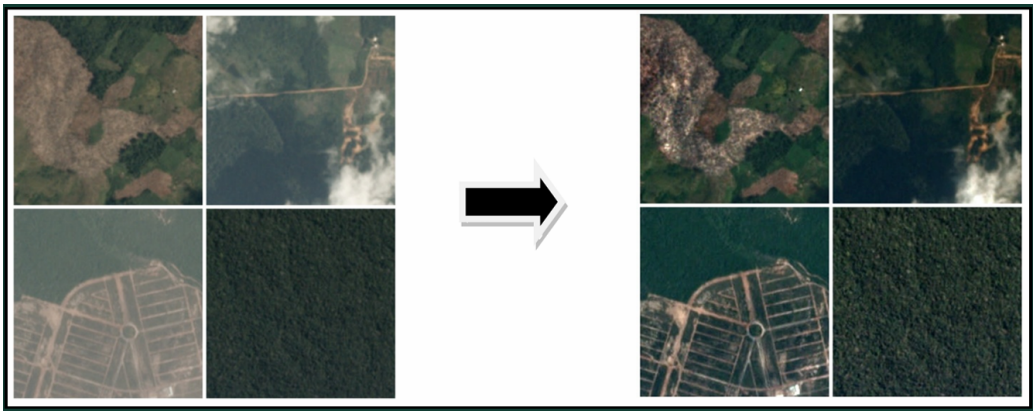

Figure 4.2: Applying haze removal on our dataset

Satellite imagery often suffers from atmospheric haze that obscures fine details. To mitigate this, we apply the Dark Channel Prior dehazing algorithm [12]. Within each small patch $\Omega(x)$ of the input image I, we compute the dark channel

$$J_{\text{dark}}(x) = \min_{y \in \Omega(x)} \Big(\min_{c \in \{r,g,b\}} I^c(y) \Big)$$

which highlights the lowest-intensity pixels across all color channels. The transmission map $t(x)$ is then estimated by

$$t(x) = 1 - \omega \min_{y \in \Omega(x)} \Big(\min_{c} \frac{I^c(y)}{A^c} \Big)$$

where A represents the global atmospheric light and $\omega = 0.95$. Finally, the haze-free image J is recovered per channel as

$$J^c(x) = \frac{I^c(x) - A^c}{\max\big(t(x),\, t_0\big)} + A^c$$

with $t_0 = 0.1$ preventing division by very small values. This dehazing step enhances contrast and reveals otherwise obscured textures critical for accurate classification as given in Fig 4.2.

4.2 Process Pipeline

Our approach follows a structured pipeline that begins with the preprocessing of satellite imagery and proceeds through stages of feature extraction and classification using a diverse set of deep learning models. These models are fine-tuned on a multi-label dataset derived from the Kaggle Planet: Understanding the Amazon from Space challenge, which provides satellite imagery tagged with environmental labels. The primary objective is to detect multiple land cover and land use attributes present in each image, such as haze, agriculture, water, and primary forest, among others.

4.2.1 CNN-Based Architectures

To establish a strong baseline, we first employed several well-established Convolutional Neural Network (CNN) architectures that have proven their effectiveness in a variety of image classification tasks. Specifically, we utilized MobileNetV2, ResNet-50, DenseNet-121, and Inception-v3, all of which were pre-trained on the ImageNet dataset. These models serve as robust feature extractors, capable of capturing both low-level and high-level visual features from the satellite images.

Among these, MobileNetV2 was particularly emphasized due to its lightweight design and efficiency in low-resource settings—a crucial factor when deploying models in real-world scenarios involving large volumes of satellite data or limited computational capacity. It leverages depthwise separable convolutions and inverted residual connections, which significantly reduce the number of parameters without compromising performance.

To adapt each CNN model for our multi-label classification task, we made the following architectural adjustments:

- Replaced the final softmax classification layer with a customized head tailored to our task.
- Added a Global Average Pooling layer to compress the spatial dimensions of feature maps.
- Appended a dense (fully connected) layer with 128 neurons and ReLU activation to introduce non-linearity.
- Incorporated a dropout layer with a rate of 0.5 to prevent overfitting.
- Finalized with a dense output layer comprising 17 units (corresponding to the 17 possible labels), each activated using the sigmoid function to allow for independent class probabilities.

These modifications enable the CNNs to handle the multi-label nature of the dataset while ensuring generalization across diverse image features.

4.2.2 Transformer-Based Architectures

To further push the boundaries of our classification performance, we integrated transformer-based models into our study—specifically, lightweight and efficient Vision Transformer (ViT) variants. These models introduce a paradigm shift from traditional convolutional methods by replacing localized receptive fields with self-attention mechanisms, which are inherently capable of capturing long-range dependencies across the image.

The transformer models explored in this study include:

- **ViT-Tiny-Patch16-224**
- **ViT-Small-Patch16-224**
- **ViT-Base-Patch16-224**
- **Swin-Tiny-Patch4-Window7-224**
- **DeiT-Tiny-Patch16-224**

- **DeiT-Base-Patch16-224**

Each of these models was initialized with ImageNet21k pre-trained weights, allowing them to benefit from rich semantic priors learned from large-scale image corpora. The baseline ViT architecture processes input images by dividing them into fixed-size patches, embedding these patches, and feeding the resulting sequence into a stack of transformer encoders. Each encoder consists of:

- A multi-head self-attention module that learns to weigh different patches based on their relevance.
- A feedforward neural network (FFN) that refines the representations learned from the attention mechanism.

This setup allows the model to flexibly focus on important image regions—such as cloud coverage, water bodies, or vegetation—regardless of their position or size, making it especially powerful for multi-label remote sensing classification.

To tailor these models for our task, we appended the following layers to their output heads:

- A dense layer with 128 units to map the high-dimensional transformer features into a compact representation.
- A dropout layer (rate = 0.5) to reduce overfitting during training.
- A final dense output layer with 17 units, each equipped with a sigmoid activation function to produce independent probability scores for each class label.

Through this dual-track architecture—comparing CNNs and transformer-based models—we aim to evaluate the trade-offs between computational cost, model complexity, and classification performance in the context of satellite imagery with multiple labels.

4.3 Training Procedure

For both CNN and transformer-based models, the same training procedure is adopted to ensure fair comparisons. We use the Adam optimizer with an initial learning rate of 10^{-4}, adjusted dynamically using a learning rate scheduler across 14 epochs. The loss function employed is the binary cross-entropy loss, formulated as:

$$\mathcal{L} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{17}\left[y_{n,i}\log(p_{n,i}) + (1-y_{n,i})\log(1-p_{n,i})\right]$$

where $y_{n,i}$ is the ground truth and $p_{n,i}$ is the predicted probability for the i-th class in the n-th sample.

4.4 Thresholding Strategy for Prediction

In multi-label classification tasks, it is essential to convert the continuous-valued probabilities—produced by the sigmoid activation in the final output layer—into discrete binary class labels that indicate the presence or absence of each category. This step is crucial for interpreting the model's output in a real-world context, such as identifying whether a satellite image contains signs of agriculture, water bodies, or deforestation.

To achieve this, we employ a thresholding strategy that maps each posterior class probability p i to a binary label ^{y}

i . Specifically, a fixed threshold of = 0.2 t=0.2 is applied across all classes. This value was not chosen arbitrarily; it was selected based on extensive empirical experimentation and validation. Multiple threshold values were evaluated to find a balance that optimally manages the trade-off between precision (minimizing false positives) and recall (minimizing false negatives), which is particularly important in environmental monitoring where missing a label could have significant consequences.

The final predicted label for each class is computed using the following rule:

$$\hat{y}_i = \begin{cases} 1, & \text{if } p_i > 0.2, \\ 0, & \text{otherwise.} \end{cases} \tag{4.2}$$

This means that if the model assigns a probability greater than 0.2 to a specific class, we interpret it as a confident prediction that the corresponding attribute is present in the image. Otherwise, the class is considered absent.

While more dynamic thresholding techniques (such as class-specific or sample-dependent thresholds) could potentially offer marginal gains, we found that using a global threshold of 0.2 provided a good balance between simplicity, interpretability, and performance on our validation set.

4.5 One-Hot Encoding and Label Co-Occurrence

The multi-label annotations are one-hot encoded during training. This approach enables the model to handle interdependent class labels such as "clear/cloudy," "water," "roads," and "urbanization," which frequently co-occur in the Amazon region. By learning from these co-occurrences, the models become more adept at detecting complex and overlapping environmental features.

4.6 Hard Fusion Strategy

In this phase, a **hard fusion strategy** was implemented to consolidate the predictions from the best-performing models identified in both Paper 1 and Paper 2. The selected models included **MobileNetV2** (from Paper 1) and **ViT_tiny_patch16_224**. These models were chosen based on their high accuracy and F-beta scores, providing a well-rounded balance between computational efficiency and predictive performance.

The fusion approach employed is known as **hard voting** or **majority rule voting**. In this ensemble method, the final label prediction for each satellite image was determined by aggregating the outputs of the individual models. Let $M_1, M_2, \ldots, M_n$ denote the n models participating in the ensemble (here, $n = 2$), and let $y_i^{(j)} \in \{0, 1\}$ be the binary prediction for label i by model M_j. Then the final prediction $\hat{y}_i$ for label i is given by:

$$\hat{y}_i = \begin{cases} 1, & \text{if } \sum_{j=1}^{n} y_i^{(j)} \geq \lceil \frac{n}{2} \rceil \\ 0, & \text{otherwise} \end{cases} \tag{4.3}$$

This strategy ensures that a label i is included in the final output only if the majority of models agree on its presence.

Using hard fusion brought several important benefits. It improved the stability of predictions, making the system less sensitive to noise and overfitting from any single model. Additionally, it enhanced the robustness of classifying less common labels in multi-label datasets. This approach was especially valuable in satellite-based environmental monitoring, where minimizing false positives is crucial to maintaining accuracy.

In summary, the hard fusion mechanism allowed the system to harness the strengths of each individual model while offsetting their weaknesses, leading to better overall classification performance and improved generalization.

Chapter 5

RESULTS and DISCUSSION

This section presents and compares the experimental results obtained from the two distinct methodologies employed in our study. The first paper investigates a CNN-based approach including models such as MobileNetV2, ResNet-50, DenseNet-121, Vision Transformer (ViT), and Inception-v3. The second paper explores various lightweight Vision Transformer architectures made available through the Timm library, including ViT-Tiny, ViT-Small, Swin-Tiny, DeiT-Tiny, and DeiT-Base.

5.0.1 Paper 1: CNN and Vision Transformer Results

The performance metrics—training accuracy, testing accuracy, and F-beta score—were used to evaluate the effectiveness of different deep learning models for multi-label classification. Table 5.1 summarizes the results from Paper 1.

Table 5.1: Performance Comparison of Models in Paper 1

Model (Training)	**Accuracy (Testing)**	**F-beta Score**
MobileNetV2	95.79%	0.9275
ResNet-50	95.56%	0.9240
DenseNet-121	95.70%	0.9248
Vision Transformer (ViT)	95.66%	0.9238
Inception-v3	93.50%	0.8920

Among the models evaluated in Paper 1, MobileNetV2 achieved the highest F-beta score of 0.9275 with a testing accuracy of 95.79%, outperforming other CNN-based models and even the baseline Vision Transformer. This validates the strength of MobileNetV2 in balancing accuracy and computational efficiency in multi-label classification tasks on satellite imagery.

5.0.2 Paper 2: Lightweight Transformer Model Results

Paper 2 focuses on fine-tuning lightweight variants of Vision Transformer architectures using the Timm library. The evaluation includes training accuracy, and F-beta scores on both training and validation sets, as shown in Table 5.2.

From Paper 2, the Swin-Tiny variant achieved the highest accuracy at 96.64%, while DeiT-Base reported the highest training F-beta score of 0.9612. However, in terms of validation F-beta score—a more generalizable metric—ViT-Tiny performed best at 0.9252, closely followed by DeiT-Tiny at 0.9242. These results highlight the effectiveness of

Table 5.2: Performance of Lightweight Transformer Models (Paper 2)

Model	**Accuracy (%)**	**Train F-beta Score**	**Validation F-beta Score**
ViT (Timm baseline)	95.66%	0.9444	0.9238
ViT_small_patch16_224	95.61%	0.9497	0.9215
ViT_base_patch16_224	95.21%	0.9301	0.9190
ViT_tiny_patch16_224	95.67%	0.9445	**0.9252**
Swin_tiny_patch4_224	**96.64%**	0.9450	0.8990
DeiT_tiny_patch16_224	95.50%	0.9485	0.9242
DeiT_base_patch16_224	95.85%	**0.9612**	0.9218

lightweight transformer models for multi-label classification tasks and suggest that ViT-Tiny strikes the best balance between performance and efficiency.

5.0.3 Comparative Summary

Both papers demonstrate strong performance on multi-label satellite image classification tasks. While the first paper shows that MobileNetV2 outperforms other CNNs and the baseline ViT in F-beta score, the second paper reveals that lightweight transformer variants, particularly ViT-Tiny and DeiT-Tiny, can match and even exceed that performance. Notably, ViT-Tiny achieved the highest validation F-beta score (0.9252) among all models across both studies. These findings suggest that transformer-based models, when appropriately scaled and fine-tuned, can serve as powerful alternatives to CNNs for environmental monitoring tasks using remote sensing imagery.

Chapter 6

LIST OF PUBLICATIONS

1. T. Barman and S. Susan, "Multi-Label Remote Sensing Image Classification using MobileNetV2," *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Kamand, India, 2024, pp. 1–4, doi: https://doi.org/10.1109/ICCCNT61001.2024.10725506.
   Keywords: Deep learning; Training; Rainforests; Satellites; Image coding; Computational modeling; Transformers; Satellite images; Remote sensing; Image classification; Multi-label image classification; Pre-trained convolutional neural networks.

2. T. Barman and S. Susan, "Multi-Label Satellite Image Classification using Lightweight Vision Transformers," *2025 IEEE 14th International Conference on Communication Systems and Network Technologies (CSNT)*, Bhopal, India, 2025, pp. 564–567, doi: https://doi.org/10.1109/CSNT64827.2025.10968132.
   Keywords: Deep learning; Computer vision; Rainforests; Computational modeling; Transformers; Feature extraction; Data models; Satellite images; Convolutional neural networks; Remote sensing; Multi-label image classification; Lightweight vision transformers.

Chapter 7

CONCLUSION AND FUTURE SCOPE

With the advent of high-resolution satellite imagery and the rapid growth of satellite imaging enterprises, there has emerged an urgent necessity for the development of accurate, efficient, and scalable methods for processing and interpreting vast volumes of remote sensing data. This research work is motivated by the need to facilitate the understanding of the complex land-use dynamics and environmental threats in the Amazon rainforest—a region rich in biodiversity, ecological importance, and subject to increasing anthropogenic pressures. To address the challenge of classifying multi-label satellite images, particularly those with heterogeneous land cover types, atmospheric conditions, and varying scales of anthropogenic activities, two complementary deep learning approaches were investigated.

The first study focused on leveraging the strengths of state-of-the-art convolutional neural networks (CNNs), including ResNet-50, DenseNet-121, Inception-v3, and MobileNetV2, along with Vision Transformer (ViT). A comprehensive evaluation of these models revealed that MobileNetV2 delivered the most effective classification performance, achieving a high testing accuracy of 95.79% and an F-beta score of 0.9275. This model's lightweight nature and efficiency make it particularly suitable for large-scale image analysis tasks in computationally constrained environments. The preprocessing pipeline in this phase incorporated haze removal using the dark channel prior technique, normalization, augmentation, and one-hot encoding of the multiple class labels, thereby ensuring that the models could effectively learn the diverse semantic features present in the Amazon satellite dataset.

Parallelly, the second study investigated the efficacy of lightweight Vision Transformer (ViT) models made available through the Timm library. The models evaluated included ViT-Small, ViT-Tiny, ViT-Base, Swin-Tiny, DeiT-Tiny, and DeiT-Base. These transformer-based models offer a powerful alternative to conventional CNNs, particularly in their ability to capture long-range dependencies and global contextual information through self-attention mechanisms. Among the evaluated models, ViT-Tiny emerged as the top performer, achieving a validation F-beta score of 0.9252 and a testing accuracy of 95.67%. DeiT-Base also showcased competitive performance with the highest training F-beta score of 0.9612, underscoring the capability of these architectures in multi-label classification tasks.

Recognizing the complementary strengths of the CNN-based and Transformer-based models, a fusion strategy was designed in the final phase of this research. Specifically, the best-performing models from both studies—MobileNetV2 and ViT-Tiny—were integrated using a hard fusion ensemble approach. In this strategy, predictions from both models were combined at the decision level using a predefined thresholding mechanism, and a majority voting scheme was applied to generate the final multi-label output. This approach is advantageous as it utilizes the feature extraction efficiency of MobileNetV2

and the global contextual understanding of ViT-Tiny, thereby enhancing the robustness and reliability of predictions.

The fusion framework involved preprocessing the input image, performing forward passes through both models to obtain respective probability vectors, and then converting these vectors into binary predictions using a common threshold (0.2). A hard voting mechanism was then applied to determine the final class labels. This ensemble not only improved classification consistency but also offered resilience against the noise and ambiguities often present in remote sensing data. The integration strategy demonstrated significant promise in producing reliable multi-label classifications across seventeen annotated class labels, encompassing land cover types, water bodies, vegetation states, and anthropogenic impacts.

Furthermore, the utility of the proposed framework extends beyond academic interest. The ability to accurately classify and monitor land-use changes in the Amazon has profound implications for environmental sustainability and policy-making. Illegal mining activities, particularly artisanal mines that are unregulated and often invisible to authorities, pose severe threats to biodiversity, water quality, and indigenous communities. The proposed classification system can serve as a crucial tool in identifying such activities from satellite images, thereby supporting government efforts in environmental protection and resource management.

The studies collectively underscore the importance of combining deep learning methodologies for enhancing performance in multi-label satellite image classification. While CNNs excel in capturing local spatial features, Vision Transformers bring the benefit of holistic scene understanding. Their combination via ensemble techniques represents a powerful paradigm for improving the accuracy and generalizability of classification frameworks in the domain of remote sensing.

Looking forward, several avenues exist to further improve this work. These include exploring more sophisticated fusion techniques such as weighted voting or trainable fusion networks, incorporating temporal information from satellite image sequences to detect dynamic changes, and utilizing additional satellite modalities such as hyperspectral and SAR data. Moreover, expanding the study to encompass other ecologically sensitive regions and integrating socio-economic datasets could provide a more comprehensive tool for global environmental monitoring.

In conclusion, this research contributes significantly to the fields of remote sensing, environmental informatics, and deep learning by developing and evaluating an efficient and accurate framework for multi-label classification of satellite imagery. By integrating the strengths of CNN and Transformer models, and leveraging hard fusion for ensemble prediction, this work sets a foundation for future advancements in automated land-use classification and deforestation monitoring. It highlights the role of machine learning in supporting sustainable development goals and offers a scalable solution for addressing environmental challenges through intelligent image analysis.

Bibliography

[1] Frederic Achard, Hugh D Eva, Hans-Jurgen Stibig, Philippe Mayaux, Javier Gallego, Thomas Richards, and Jean-Paul Malingreau. Determination of deforestation rates of the world's humid tropical forests. *Science*, 297(5583):999–1002, 2002.

[2] John E Ball, Derek T Anderson, and Chee Seng Chan. A comprehensive survey of deep learning in remote sensing: theories, tools and challenges for the community. *IEEE Geoscience and Remote Sensing Magazine*, 5(4):8–36, 2017.

[3] Tamal Barman and Seba Susan. Multi-label remote sensing image classification using mobilenetv2. 2024.

[4] P. Coppin and M. Bauer. Digital change detection in forest ecosystems with remote sensing imagery: A review. *Remote Sensing Reviews*, 13(3-4):207–234, 2014.

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[7] Matthias Drusch, Umberto Del Bello, Serge Carlier, Olivier Colin, Victor Fernandez, Ferran Gascon, Bernhard Hoersch, Claudio Isola, Paolo Laberinti, Philippe Martimort, et al. Sentinel-2: Esa's optical high-resolution mission for gmes operational services. *Remote Sensing of Environment*, 120:25–36, 2024.

[8] Giles M Foody. Status of land cover classification accuracy assessment. *Remote sensing of environment*, 80(1):185–201, 2002.

[9] Holly K Gibbs et al. Mapping and understanding land-use frontiers in the global tropics. *Environmental Research Letters*, 10(12):125002, 2015.

[10] Scott J Goetz, David Steinberg, Matthew G Betts, Richard T Holmes, Patrick J Doran, and Ralph Dubayah. Remote sensing of biodiversity: prospects and constraints. *Ecological Applications*, 15(4):1177–1186, 2015.

[11] Matthew C Hansen, Peter V Potapov, Rebecca Moore, Michael Hancher, Svetlana A Turubanova, Alexandra Tyukavina, David Thau, Stephen V Stehman, Scott J Goetz, Thomas R Loveland, et al. High-resolution global maps of 21st-century forest cover change. *Science*, 342(6160):850–853, 2013.

[12] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12):2341–2353, 2011.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 4700–4708, 2017.

[15] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[16] Shaohui Ji, Shaohui Wei, Meng Lu, Zhen Wang, and Lin Li. A survey on knowledge transfer for deep learning in remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing*, 164:61–84, 2020.

[17] Christopher O Justice, Eric Vermote, John RG Townshend, Ruth Defries, David P Roy, Dorothy K Hall, Vincent V Salomonson, Jeffrey L Privette, George Riggs, Alan Strahler, et al. An overview of modis land data processing and product status. *Remote sensing of Environment*, 83(1-2):3–15, 2002.

[18] Maria Kaselimi, Anastasios Doulamis, Nikolaos Doulamis, Elias Kalogerakis, Dimitrios Zarpalas, and Petros Daras. Transformers in remote sensing: A review and outlook. *ISPRS Journal of Photogrammetry and Remote Sensing*, 200:202–222, 2023.

[19] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM Computing Surveys (CSUR)*, 54(10s):1–41, 2022.

[20] Dengsheng Lu and Qihao Weng. Remote sensing of forest health: A review. *Remote sensing of environment*, 112(5):2995–3006, 2016.

[21] Natalie M Mahowald et al. The impact of land use change on climate and carbon cycle. *Nature Geoscience*, 10(10):758–765, 2017.

[22] J. Rangel et al. A survey on convolutional neural networks and their performance limitations in image recognition tasks. *Journal of Sensors*, 2024:2797320, 2024.

[23] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.

[24] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016.

[25] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2020.

[26] United Nations Framework Convention on Climate Change (UNFCCC). Redd+ - reducing emissions from deforestation and forest degradation, 2021. Accessed: 2025-05-21.

[27] Di Wang, Qiming Zhang, Yufei Xu, Jing Zhang, Bo Du, Dacheng Tao, and Liangpei Zhang. Advancing plain vision transformer towards remote sensing foundation model. *arXiv preprint arXiv:2208.03987*, 2022.

[28] Yi Wang, Conrad M Albrecht, and Xiao Xiang Zhu. Self-supervised vision transformers for joint sar-optical representation learning. *arXiv preprint arXiv:2204.05381*, 2022.

[29] Curtis E Woodcock, Rebecca Allen, Matt Anderson, Alan Belward, Robert Bindschadler, Warren Cohen, Feng Gao, Samuel Goward, Dennis Helder, Eileen Helmer, et al. Free access to landsat imagery. *Science*, 320(5879):1011–1011, 2008.

[30] Zhi Yao, Xueyang Gao, Jianya Wang, Wei Li, and Yan Zhang. Land-cover classification using vision transformers on sentinel-2 imagery. *Remote Sensing*, 13(14):2755, 2021.

[31] Xiao Xiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 5(4):8–36, 2017.