Online Version Available at: https://jurnal.ubd.ac.id/index.php/te

JOURNAL OF TECH-E

| 2581-1916 (ONLINE) | 2598-7585 (PRINTED) |

Article

Optimization of CNN and Vision Transformer Models in Addressing Long-Tailed Data Imbalance for Satellite Cloud Image Classification

Revatta Manggala Nandivadhano[1], *Aditiya Hermawan*[2], *Lidya Lunardi*[3]

[1, 2, 3]*Buddhi Dharma University, Informatics Engineering, Banten, Indonesia*

SUBMISSION TRACK

Recieved: 11, 25, 2025
Final Revision: 01, 05, 2026
Available Online: 02, 02, 2026

KEYWORD

Long-Tailed Learning, Vision Transformer, RemoteCLIP, Satellite Cloud Classification, Class-Balanced Loss

CORRESPONDENCE

E-mail:
manggala.nandi@gmail.com
aditiya.hermawan@ubd.ac.id
lidya.lunardi@ubd.ac.id

A B S T R A C T

This study investigates long-tailed satellite cloud image classification by comparing CNN and Vision Transformers (ViT) built upon vision–language foundation models. A large-scale satellite cloud dataset with 11 highly imbalanced classes, including a dominant non-phenomenon category, is used to represent realistic atmospheric variability. The data are split using stratified sampling, standardized to a fixed resolution, and used to fine-tune CLIP-based backbones from RemoteCLIP and GeoRSCLIP through parameter-efficient adaptation. Several loss functions Cross Entropy, Logit Adjustment, Focal, Class-Balanced, and label-distribution–aware variants are evaluated, along with experiments examining majority-class removal and adapter bottleneck adjustments. Initial results show that Logit Adjustment causes majority-class collapse under default settings. After optimization, ViT-based models consistently outperform CNN models, achieving higher accuracy and more balanced macro-level performance. Class-Balanced loss emerges as the most effective objective, offering a strong trade-off between overall accuracy and per-class fairness. Increasing the adapter bottleneck dimension further boosts ViT performance, enabling the best configuration to match or exceed prior benchmarks while improving minority-class recognition. The final optimized model is deployed in a web-based prediction system, demonstrating the practical potential of foundation-model approaches for satellite-driven weather analysis.

DOI: 10.31253/te.v9i2.4256

I. INTRODUCTION

Weather plays a fundamental role in human life, as extreme weather conditions can disrupt daily activities and, in the long term, stimulate climate-related migration [1]. The frequency and intensity of extreme weather events continue to increase along with the worsening impacts of climate change [2]. This ongoing anthropogenic climate change is primarily driven by the rise of greenhouse gas emissions released into the atmosphere [3], As climate conditions continue to shift and alter the characteristics of weather phenomena, the ability to detect atmospheric patterns accurately and continuously has become increasingly important. One promising approach for inferring weather systems is through the analysis of cloud structures derived from satellite imagery [4].

Meteorological satellites produce cloud images that contain essential information for weather monitoring, climate analysis, and early-warning systems. Such images have been widely used as an alternative to conventional weather detection methods [5]. The process of extracting information from these satellite images falls within the broader field of remote sensing, which involves observation of the Earth from a distance and often incorporates machine learning techniques to analyze large-scale environmental data [6], [7], [8], [9], [10]. In recent years, machine learning—and particularly deep learning has shown considerable potential in meteorology-related remote sensing tasks. Convolutional Neural Networks (CNNs) have long served as the dominant architecture for image classification, target detection, and semantic segmentation due to their ability to capture spatial and spectral patterns [11], [12], [13]. Meanwhile, the Vision Transformer (ViT) has emerged as a strong alternative model that applies Transformer mechanisms to visual data by computing self-attention across image patches [14]. The self-attention structure enables ViT to model both local and global relationships within images [15], and several studies report that ViT surpasses CNN performance in various computer vision tasks [16]. However, other research indicates that CNNs still excel in certain classification scenarios [12], suggesting that the comparative effectiveness of CNN and ViT remains dependent on data characteristics and task complexity. These observations highlight the need to further evaluate both architectures for satellite cloud classification.

A widely used benchmark dataset for cloud classification is the Large-Scale Satellite Cloud Image Database for Meteorological Research (LSCIDMR) [4]. Although effective, the dataset poses challenges due to inter-class similarity and severe class imbalance [17]. Some previous studies addressed the imbalance by excluding the extremely imbalanced LabelLess class, which was found to significantly degrade model quality [17]. Other studies chose to retain this class and still achieved strong results [5]. Including the LabelLess class is important because it captures cloud scenes without dominant weather phenomena, thus providing broader representativeness of atmospheric conditions. Nevertheless, the presence of class imbalance requires specialized techniques to prevent the model from being biased toward majority classes.

To overcome the challenges of long-tailed datasets, recent studies introduced advanced fine-tuning methods such as Long-Tail Learning with Foundation Models (LIFT), which allows structured and lightweight adaptation of large pretrained models to imbalanced data [18]. The use of foundation models particularly domain-specific visual-language models such as RemoteCLIP [19] and GeoRSCLIP [13] provides additional potential for improving performance in remote sensing tasks by leveraging knowledge learned from large-scale pretraining. This study builds upon these developments to assess the capabilities of CNN and ViT when applied to a large, long-tailed satellite cloud dataset.

Beyond the technical challenges, the classification of cloud imagery plays a significant role in practical meteorology. Cloud morphology contains valuable cues about atmospheric dynamics, such as the presence of convective systems, storm development, or the transition between different atmospheric states. Features such as texture, spatial distribution, brightness gradients, and cloud-top height all provide important information for operational forecasting. However, these features can

DOI: 10.31253/te.v9i2.4256

vary widely even within the same class, leading to substantial intra-class variance, while visually similar cloud types from different categories can introduce inter-class ambiguity. These characteristics make satellite cloud classification a highly complex problem, requiring models that can generalize across diverse atmospheric patterns while remaining robust to imbalance and noise.

From a modeling perspective, CNNs have traditionally excelled at identifying localized spatial features due to their convolution-based structure. Operations such as convolution and pooling allow CNNs to capture fine-grained textures, edge structures, and repetitive patterns that often appear in cloud imagery. This property is advantageous for identifying phenomena like cirrus streaks, cumuliform structures, or regions of dense convection. Nonetheless, CNNs depend on hierarchical feature extraction to expand their receptive field, which can limit their ability to capture long-range dependencies and global cloud structures without significantly increasing model depth. In contrast, ViT mechanisms enable direct modeling of global relationships through multi-head self-attention. This structure allows the model to learn spatial dependencies across distant regions of the image, making it particularly effective for recognizing large-scale cloud formations such as typhoons, frontal systems, or mesoscale convective complexes. Global modeling is essential in meteorology because many atmospheric phenomena are not defined solely by local features but by broader spatial coherence. For this reason, ViT-based architectures are increasingly explored for remote sensing data, though their performance under severe class imbalance remains insufficiently studied.

Furthermore, the use of vision–language foundation models introduces additional advantages. Models such as RemoteCLIP and GeoRSCLIP encode semantic relationships between visual features and textual descriptions, enabling better generalization when fine-tuned on limited or imbalanced datasets. Parameter-efficient tuning approaches such as adapters or prompt-based learning make it possible to adapt these large models without excessive computational cost. Despite the promise of these methods, the extent to which CNN and ViT architectures benefit from foundation model pretraining especially under a heavily long-tailed label distribution remains an open question that this research aims to address.

The objective of this research is to model and compare CNN and ViT for long-tailed satellite cloud image classification using the LSCIDMR dataset. This includes analyzing their performance differences, evaluating their suitability for weather system inference, and exploring optimization strategies to handle severe class imbalance. The study also aims to implement fine-tuning approaches that enhance model robustness in long-tailed scenarios and develop a web-based application that deploys the best-performing model for practical meteorological use. Through this approach, the research contributes to the improvement of satellite-based weather inference by identifying the strengths of both architectures and determining effective optimization techniques for imbalanced cloud classification tasks.

II. LITERATURES REVIEW

Deep learning has become a dominant approach in remote sensing, particularly for analyzing satellite and UAV imagery through convolutional neural networks (CNNs) and Vision Transformers (ViTs). A key challenge in this field is that many remote sensing datasets exhibit long-tailed label distributions, which require specialized strategies to prevent performance degradation on minority classes. One of the most widely used datasets for satellite cloud classification is LSCIDMR, which contains over 100,000 high-resolution images with single-label and multi-label annotations [4]. Benchmark results show that EfficientNet performs best on single-label configurations, while EfficientNet and AlexNet lead on multi-label tasks. Subsequent studies have proposed various optimization strategies on LSCIDMR. A recent study introduced CDC-Net, a lightweight CNN that applies depth-wise convolutions, FeatureCopy operations, and local importance-based pooling, achieving faster inference and higher accuracy than previous lightweight models [5]. A snapshot-based residual network (SnapResNet) was introduced, achieving 97.25% accuracy on the 10-class version of LSCIDMR-S, although its performance deteriorated when

DOI: 10.31253/te.v9i2.4256

applied to the full 11-class configuration, highlighting the difficulty of modeling the dataset's long-tailed distribution [17].

Beyond CNN-based methods, Transformers have been explored as alternative backbones for remote sensing tasks due to their ability to capture long-range dependencies. Previous work showed that Transformers are not inherently more robust than CNNs under adversarial attacks, but they tend to outperform CNNs on out-of-distribution samples when training conditions are standardized, suggesting an architectural advantage [4]. Research on long-tailed learning further indicates that extensive fine-tuning of foundation models can degrade performance on tail classes; one study addressed this issue by proposing LIFT, a lightweight fine-tuning approach that reduces training cost while improving predictive performance [18]. ViT-based models have shown strong potential in remote sensing applications. A recent study demonstrated that ViT can outperform EfficientNet and ResNet in weed and crop classification from UAV imagery, even with limited training data [20]. A study developed ForestViT for multilabel deforestation monitoring and reported superior performance on minority classes within imbalanced datasets [15]. For multispectral imagery, another work designed a **V**iT-based U-shaped architecture with Transformer Sub-Pixel blocks for 44-class land-cover classification, achieving high accuracy by effectively modeling spatial–spectral dependencies [16].

Recent advancements have extended remote sensing into the domain of vision–language foundation models. RemoteCLIP was introduced as the first vision–language foundation model tailored for remote sensing by converting heterogeneous annotated datasets into image–caption pairs [19]. RemoteCLIP, using ResNet-50, ViT-B-32, and ViT-L-14 encoders, consistently outperforms base CLIP models across multiple tasks. Similarly, a large-scale remote sensing image–text dataset containing five million samples was constructed and used to develop a domain-pretrained CLIP variant, which demonstrated substantial performance gains across numerous downstream benchmarks [21]. CNNs remain relevant in other remote sensing applications, such as roof and material classification [22], water-body segmentation using multi-feature extraction modules [13], and anomaly detection in photovoltaic power plants with region-based CNNs that achieve high true-positive rates and low false positives [23]. Transformer-based approaches have also been adapted for meteorological cloud recognition, as demonstrated by UATNet [24], which integrates modified Swin Transformer blocks and attention-based fusion mechanisms to achieve strong performance on the CRMSCD dataset. Meanwhile, one study reported that deep CNNs outperform ViTs for single-stage tropical cyclone intensity estimation, although hybrid ViT–CNN approaches can provide slight improvements, suggesting that CNNs may remain preferable for certain meteorological tasks [25].

Overall, existing studies demonstrate that both CNNs and ViTs are powerful architectures for remote sensing, yet their performance can vary across tasks, modalities, and dataset characteristics. Despite extensive work on LSCIDMR, most research is still limited to CNN-based models or reduced, more balanced subsets of the dataset, often excluding additional classes that contribute to the long-tailed nature of the full dataset. It has been noted that incorporating the complete label set can result in suboptimal performance, and therefore remains insufficiently explored [17]. Consequently, there is no comprehensive comparison of CNN and ViT foundation models under the full long-tailed LSCIDMR setting, nor an evaluation of parameter-efficient fine-tuning methods to mitigate performance degradation on minority classes. Addressing this gap, the present study compares CNN- and ViT-based foundation models on the long-tailed satellite cloud imagery of LSCIDMR using the LIFT method, with the aim of assessing their relative performance and determining whether ViT-based models can consistently outperform CNNs when both are fine-tuned under an efficient long-tailed learning framework [18].

DOI: 10.31253/te.v9i2.4256

III. FRAMEWORK

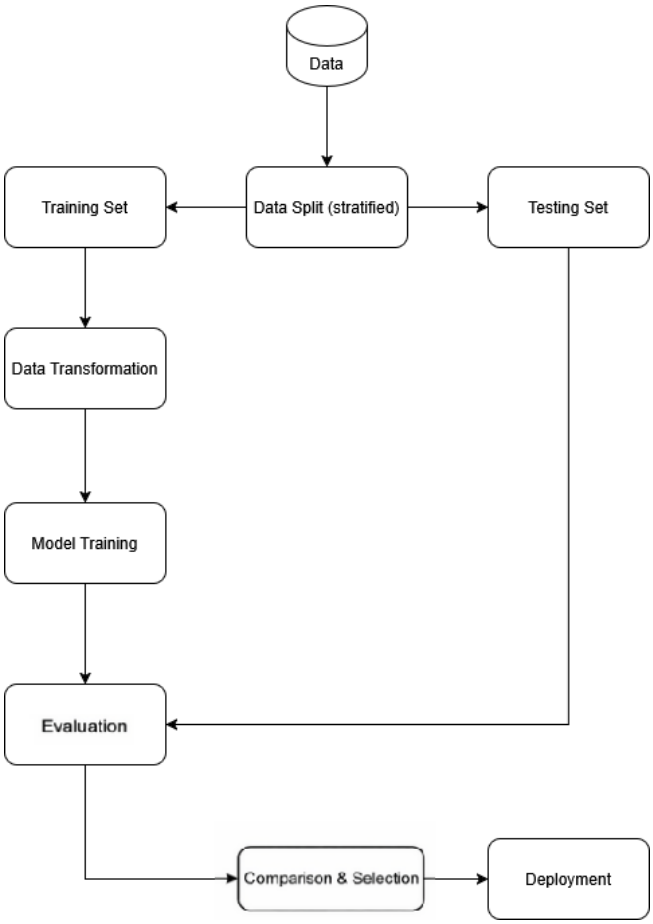


Figure 1. Framework

Figure 1 presents the workflow of the study. Given the imbalance within the dataset, a stratified data-splitting technique is employed to ensure that both subsets remain representative of the overall population. The training subset undergoes a series of simple transformations to align with the required input dimensions of the model. Once transformed, the data proceed to the training stage, during which the model is fitted and begins learning the inherent features in the dataset.

The CLIP architecture utilized in this study consists of a text encoder and an image encoder. Following the LIFT methodology, the text encoder is used to initialize the classifier's weights, while the image encoder functions as the backbone of the model. After training is completed, the model is evaluated using the test set, and its performance is compared against models employing ViT- or CNN-based backbones in order to determine which achieves superior results according to the predetermined evaluation metrics. Upon completion of the entire training pipeline, the best-performing model is selected for implementation.

IV. METHOD

This study employs a secondary dataset sourced from the repository "https://github.com/Zjut-MultimediaPlus/LSCIDMR," using the Large-Scale Satellite Cloud Image Database for Meteorological Research – Single-Label Annotation (LSCIDMR-S), provided by its original developers [4]. The dataset consists of 104,390 satellite image slices categorized into 11 classes and sourced from the Himawari-8 geostationary satellite operated through the Japan Aerospace Exploration Agency (JAXA) P-tree System. Prior to public release, the dataset creators conducted several preprocessing steps to ensure consistent quality. First, channel selection was performed by choosing three spectral channels *albedo_3*, *albedo_4*, and *albedo_5* because these channels provide color variations suitable for distinguishing between water clouds and ice clouds, as well as

DOI: 10.31253/te.v9i2.4256

differentiating ocean surfaces, vegetation, and deserts. Second, time selection was conducted by capturing images at 00:20 UTC+0, the moment when the subsolar point is centered on the composite map, yielding the clearest satellite imagery. Third, position determination restricted the dataset to the Northern Hemisphere due to its higher relevance for meteorological research. Fourth, slice size determination was applied by partitioning the original full-disk satellite images into smaller slices to reduce computational complexity during model training. Finally, slicing naming followed a structured format *YYYYMMDD_ii_jj.png*, where the components represent the date, latitudinal slice index, and longitudinal slice index. All labels corresponding to each image slice are stored in a CSV file that maps filenames to their respective class annotations.

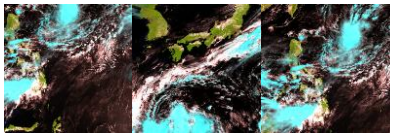
LSCIDMR-S contains 11 label classes, each representing a distinct meteorological or surface phenomenon. *Tropical Cyclone* refers to low-pressure systems featuring counterclockwise spiraling cloud structures, with inclusion determined by the presence or absence of an identifiable cyclone eye. *Extratropical Cyclone* represents elliptical low-pressure systems in mid to high latitudes, also determined by identifiable cyclone-eye features. *Frontal Surface* captures air mass boundaries characterized by cyclonic curvature and north–south cloud bands. *Westerly Jet* identifies strong upper-tropospheric jet streams visible as thin, intense west-to-east wind bands. *Snow* is recognized by dark blue imagery indicating surface-level ice crystal accumulation. *High Ice Cloud* includes cirrus, cirrostratus, and other high-altitude clouds composed primarily of ice crystals, identified through blue-tinted regions with inclusion criteria of Area(High Ice Cloud)>50% and Area(Else)<20%. *Low Water Cloud* refers to low-altitude water-droplet clouds appearing pinkish, following similar area-based inclusion thresholds. *Ocean* is represented by predominantly black regions covering more than 80% of the slice. *Desert* appears in varying shades of brown and is included when desert-like areas exceed 50% of the slice. *Vegetation* includes green-tinted regions representing forests, grasslands, or wetlands with a majority-area threshold. Lastly, *LabelLess* consists of slices that do not meet any inclusion criteria of the ten defined classes, meaning no dominant meteorological system is present.

The distribution of all 11 label classes within the dataset is presented in Table 1, providing an overview of class imbalance characteristics essential for model design. Meanwhile, representative examples of the image slices used in this study are illustrated in Table 2 to give a visual understanding of the data utilized for model training and evaluation.

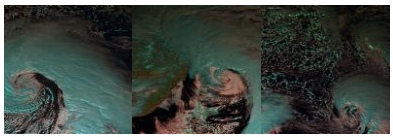
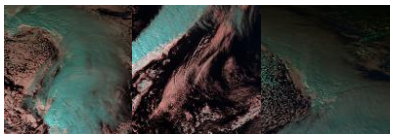
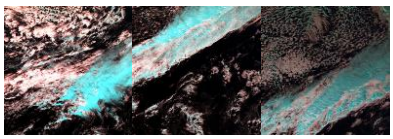
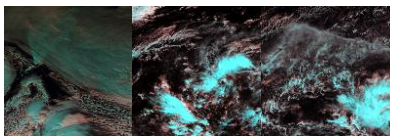
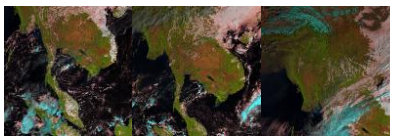
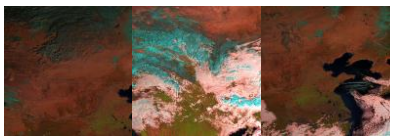
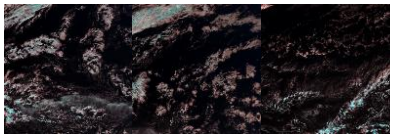
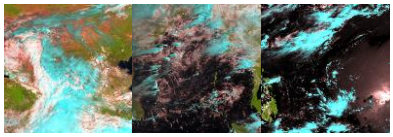
Table 1. Distribution of LSCIDMR-S Labels

Class	Count	Ratio
Tropical Cyclone	3905	3,74%
Extratropical Cyclone	4984	4,77%
Frontal Surface	634	0,607%
Westerly Jet	628	0,602%
Snow	7631	7,31%
Low Water Cloud	1774	1,70%
High Ice Cloud	5278	5,06%
Vegetation	7831	7,51%
Desert	4518	4,33%
Ocean	4042	3,88%
LabelLess	63765	61,07%
Jumlah	104390	100%

Table 2. Sample Data of LSCIDMR-S

Class	Image Sample
Tropical Cyclone	

DOI: 10.31253/te.v9i2.4256

Class	Image Sample
Extratropical Cyclone	
Frontal Surface	
Westerly Jet	
Snow	
Low Water Cloud	
High Ice Cloud	
Vegetation	
Desert	
Ocean	
LabelLess	

In the data splitting process, the dataset is divided into training and test sets using an 80:20 ratio, where 80% of the data is allocated for model training and the remaining 20% is used to evaluate model performance. The 80:20 split is one of the most commonly applied proportions in machine learning experiments [26]. The splitting procedure is conducted using a stratified approach, meaning the dataset is separated into several strata or subgroups based on class labels, and random samples are drawn from each subgroup in proportions that reflect their representation in the overall

DOI: 10.31253/te.v9i2.4256

population. Stratified sampling offers the advantage of producing samples that more accurately represent the population and enhancing the external validity of studies employing this method [27]. Given the long-tailed nature of the dataset, applying stratified sampling during data splitting is essential to ensure that minority classes remain adequately represented in both the training and test sets, while preserving the original long-tail distribution across both subsets.

The initial preprocessing stage involves resizing the input images to match the required model input dimension of 224 × 224 pixels. This transformation is performed by downscaling the images from an intermediate size of 256 × 256 pixels rather than directly from the original 1000 × 1000 resolution to preserve visual quality and reduce distortion. As illustrated in Figure 2, the original image prior to resizing maintains a larger spatial resolution, while Figure 3 shows the resulting image after the resizing process has been applied. This resizing step ensures uniformity across the dataset and prepares the images for subsequent feature extraction during model training.

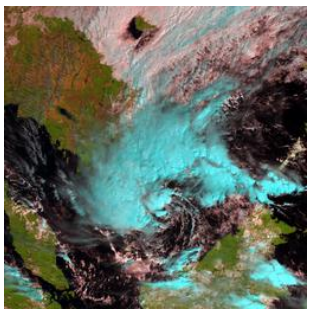


Figure 2. Image Before Resizing

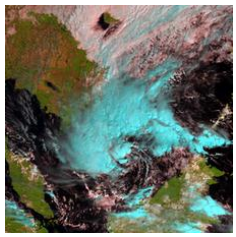


Figure 3. Image After Resizing

CLIP Architecture

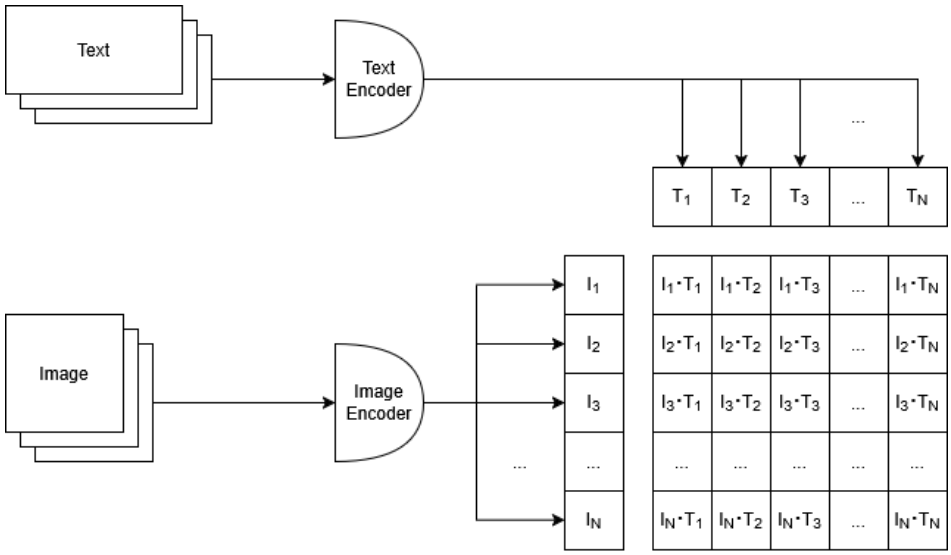


Figure 4. CLIP Contrastive Pre-training Process
Source: [28]

CLIP, or Contrastive Image–Language Pre-training, is an artificial neural network model trained on the task of predicting image captions using a contrastive learning approach [28]. CLIP employs a dual-encoder architecture consisting of a text encoder and an image encoder, where the text encoder is implemented using a Transformer. In the LIFT method utilized in this study, the CLIP text encoder is used solely for initializing the classifier's weights by extracting semantic features from the prompts, after which it is no longer employed. The prompts in this research are based on the zero-shot experiment template from RemoteCLIP, formulated as "a satellite image showing {class name}," with class names manually specified according to Table 3. Meanwhile, the

DOI: 10.31253/te.v9i2.4256

image encoder of CLIP serves as the backbone of the neural network responsible for learning features from the input data.

Table 3. Class Name in the Prompt

Label	Class Name
Desert	Desert
Extratropical Cyclone	Extratropical cyclone
Frontal Surface	Frontal surface
High Ice Cloud	High-altitude ice cloud
Low Water Cloud	Low-altitude water cloud
Ocean	Ocean
LabelLess	No dominating weather system or geographical feature
Snow	Snow cover
Tropical Cyclone	Tropical cyclone
Vegetation	Vegetation
Westerly Jet	Westerly jet stream

The hyperparameters used throughout the training process adhere to the default configuration recommended in the main results of LIFT. This choice ensures consistency with the original framework and provides a reliable baseline for comparing model performance across different architectural and loss-function settings. The specific hyperparameter values applied in this study are presented in Table 4.

Table 4. Training Hyperparameters and Their Values

Hyperparameters	Value
PEFT Method	Adaptformer (ViT) / Bias Tuning & SSF (CNN)
Epoch	10
Batch Size	128
Loss Function	Logit Adjustment
Momentum	0.9
Learning Rate	0.01
Input Size	224*224
Weight Decay	0.0005
Classifier Initialisation	Text Features

The hardware specifications used during the model training process in this study are presented in Table 5. These specifications define the computational environment in which all experiments were conducted, ensuring consistency and reproducibility across the evaluation pipeline. The details provided in the table also help contextualize model training times and performance outcomes, particularly given the computational demands of foundation-model–based architectures.

Table 5. Hardware Specifications

Hardware	Specifications
CPU	AMD Ryzen 5 8600G
GPU	NVIDIA GeForce GTX 1660 Super
RAM	16GB DDR55 4.8 GHz
Disk	512 GB SSD + 1TB HDD

The model training process was conducted within the Visual Studio Code environment using the Python programming language. Several Python libraries were utilized, including PyTorch for various machine learning tasks and, in particular, the torchvision package, which focuses on computer vision operations. Additionally, the OpenCLIP library was employed as an open-source implementation of CLIP, through which both RemoteCLIP and GeoRSCLIP used in this study are provided in the OpenCLIP format.

In this study, CLIP models from RemoteCLIP and GeoRSCLIP will be utilized [19], [21]. RemoteCLIP is available in ResNet-50, ViT-Base-32, and ViT-Large-14 variants, while

DOI: 10.31253/te.v9i2.4256

GeoRSCLIP is available in ViT-Base-32, ViT-Large-14, and ViT-Large-14 with a 336 × 336 input size. This research will focus on the ViT-Base-32 configurations from both CLIP models for the ViT-based backbone, as well as the ResNet-50 RemoteCLIP model for the CNN-based backbone. All models accept inputs with a resolution of 224 × 224.

The evaluation of model performance in this study was carried out using a Confusion Matrix as the primary assessment tool. The Confusion Matrix provides detailed information about correct and incorrect classifications across all cloud categories, making it suitable for analyzing behavior under long-tailed label distributions. The numerical values derived from this matrix were subsequently used to compute multiple performance metrics, including precision, recall, accuracy, F1-score, and macro F1-score, which together offer a comprehensive understanding of both overall performance and class-level balance.

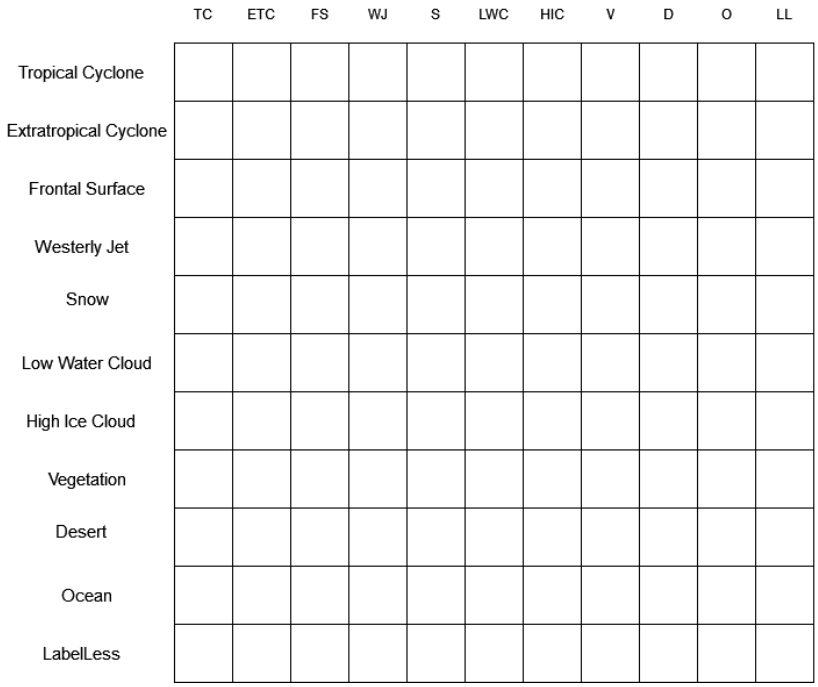


Figure 5. Confusion Matrix Illustration Using the LSCIDMR-S Dataset

V. RESULT

Table 6 presents a consolidated overview of all model evaluations performed in this study, encompassing different backbones, loss functions, and parameter-efficient configurations. The results demonstrate clear performance differences between CNN- and ViT-based architectures as well as notable shifts in model behavior depending on loss function selection. The default Logit Adjustment settings consistently produced majority collapse across all backbones, especially in ViT-B-32 and GeoRSCLIP, resulting in low accuracy and calibration issues. In contrast, Cross Entropy substantially improved overall accuracy but caused minority collapse, showing its limitations in handling long-tailed distributions. Class-Balanced loss delivered the most stable performance, offering a stronger balance between overall accuracy and per-class fairness, particularly on ViT-B-32 models. Increasing the bottleneck dimension improved performance across metrics, although with diminishing returns beyond 64 dimensions. Removing the LabelLess class produced significant accuracy gains, aligning the model more closely with earlier benchmark results. Finally, extending training to 15 epochs provided minor but consistent improvements in both total and per-class accuracy. Collectively, the table highlights how architectural choices, loss functions, and fine-tuning strategies interact to influence model robustness in long-tailed satellite cloud classification.

DOI: 10.31253/te.v9i2.4256

Table 6. Unified Evaluation Table for All Models (Models 1–18)

No	Backbone	Loss Function	Special Setting	Total	Correct	Accuracy	Macro F1	Macro Precision	Mean	ECE	Avg. Conf.	Conf (Correct	Conf. (Incorre
1	ViT-B-32	LA	Default	20878	12127	58.1%	54.4%	46.7%	83.9%	0.1385	0.7146	0.7646	0.6569
2	ViT-B-32 (GeoRSCLIP)	LA	Default	20878	11974	57.4%	54.2%	46.7%	83.9%	0.1385	0.7146	0.7646	0.6569
3	ResNet-50	LA	Default	20878	11112	53.2%	50.3%	46.7%	83.9%	0.1385	0.7146	0.7646	0.6569
4	ResNet-50	CE	-	20878	16585	79.4%	62.2%	46.7%	83.9%	0.1385	0.7146	0.7646	0.6569
5	ViT-B-32	CE	-	20878	17133	82.1%	67.7%	46.7%	83.9%	0.1385	0.7146	0.7646	0.6569
6	ViT-B-32	CB	-	20878	15616	74.8%	82.4%	46.7%	83.9%	0.1385	0.7146	0.7646	0.6569
7	ResNet-50	CB	-	20878	15298	73.3%	62.5%	46.7%	83.9%	0.1385	0.7146	0.7646	0.6569
8	ViT-B-32	LA (τ/2)	Reduced τ	20878	15713	75.3%	65.9%	46.7%	83.9%	0.1385	0.7146	0.7646	0.6569
9	ViT-B-32	Focal	-	20878	16996	81.4%	69.1%	46.7%	83.9%	0.1385	0.7146	0.7646	0.6569
10	ResNet-50	LADE	-	20878	11621	55.7%	51.6%	46.7%	83.9%	0.1385	0.7146	0.7646	0.6569
11	ViT-B-32	LA	No LabelLess	8126	7360	90.6%	85.8%	46.7%	83.9%	0.1385	0.7146	0.7646	0.6569
12	ViT-B-32	CB	No LabelLess	8126	7409	91.2%	87.6%	46.7%	83.9%	0.1385	0.7146	0.7646	0.6569
13	ViT-B-32	LA	64 Bottleneck	20878	13057	62.5%	58.0%	46.7%	83.9%	0.1385	0.7146	0.7646	0.6569
14	ViT-B-32	LA	128 Bottleneck	20878	13300	63.7%	58.8%	46.7%	83.9%	0.1385	0.7146	0.7646	0.6569
15	ViT-B-32	CB	128 Bottleneck	20878	16399	78.5%	70.6%	46.7%	83.9%	0.1385	0.7146	0.7646	0.6569
16	ViT-B-32	CE	128 Bottleneck	20878	17585	84.2%	73.0%	46.7%	83.9%	0.1385	0.7146	0.7646	0.6569
17	ViT-B-32	CB	128 Bottleneck, No LabelLess	8126	7604	93.6%	91.1%	46.7%	83.9%	0.1385	0.7146	0.7646	0.6569
18	ViT-B-32	CB	128 Bottleneck, 15 Epochs	20878	16587	79.4%	71.8%	46.7%	83.9%	0.1385	0.7146	0.7646	0.6569

DOI: 10.31253/te.v9i2.4256

Table 7 presents a consolidated comparison of all evaluated models, highlighting how architectural choices, loss functions, and training configurations influence performance. Among the initial models, the RemoteCLIP ViT-B-32 and GeoRSCLIP variants show similar results, although overall accuracy remains low due to majority-class collapse. Models trained with alternative loss functions exhibit substantial improvements, with ViT CE and ViT Focal achieving higher accuracy, while Class-Balanced loss provides stronger per-class balance as reflected in higher Mean and F1 scores. When the LabelLess class is removed, model performance increases markedly, with ViT CB 128 Dim. achieving the highest overall metrics across the full comparison. Experiments using extended parameters further demonstrate that increasing bottleneck dimensions and training epochs enhances model stability and accuracy, particularly for ViT-based models. Overall, the comparison shows that ViT architectures paired with Class-Balanced loss and extended parameter configurations consistently yield the strongest and most balanced performance.

Table 7. Summary of Model Comparison Results

No.	Category	Model	Accuracy	Mean	F1
1	Initial Models	RemoteCLIP ViT-B-32	58.1%	83.9%	54.4%
2	Initial Models	RemoteCLIP RN-50	53.2%	80.0%	50.3%
3	Initial Models	GeoRSCLIP ViT-B-32	57.4%	84.3%	54.2%
4	Other Loss Functions	RN CE	79.4%	56.4%	62.2%
5	Other Loss Functions	ViT CE	82.1%	62.8%	67.7%
6	Other Loss Functions	RN CB	73.3%	75.8%	62.5%
7	Other Loss Functions	ViT CB	74.8%	82.4%	65.5%
8	Other Loss Functions	ViT ½τ LA	75.3%	79.8%	65.9%
9	Other Loss Functions	ViT Focal	81.4%	65.2%	69.1%
10	Other Loss Functions	RN LADE	55.7%	79.1%	51.6%
11	Without LabelLess Class	ViT LA	90.6%	90.4%	85.8%
12	Without LabelLess Class	ViT CB	91.2%	90.7%	87.6%
13	Without LabelLess Class	ViT CB 128 Dim.	93.6%	93.2%	91.1%
14	Extended Parameters	64 Dim.	62.5%	86.7%	58.0%
15	Extended Parameters	128 Dim.	63.7%	86.9%	58.8%
16	Extended Parameters	CB 128	78.5%	85.0%	70.6%
17	Extended Parameters	CE 128	84.2%	68.7%	73.0%
18	Extended Parameters	CB 128 15e	79.4%	85.8%	71.8%

VI. DISCUSSION

Based on the consolidated evaluation results, it is clear that the default LIFT configuration using Logit Adjustment produces the weakest performance across all tested architectures. The initial models RemoteCLIP ViT-B-32, GeoRSCLIP ViT-B-32, and RemoteCLIP ResNet-50 achieve low accuracies of 58.1%, 57.4%, and 53.2%, respectively, with consistently poor per-class performance. The LabelLess class dominates predictions under the 60:40 imbalance ratio, reflecting Logit Adjustment's tendency to overemphasize frequent classes. Because GeoRSCLIP performs no better than RemoteCLIP and provides only marginal improvements in class averages, it is excluded from further experiments.

Replacing Logit Adjustment with Cross Entropy substantially improves performance for both CNN and ViT architectures, with accuracies rising to 79.4% for ResNet-50 and 82.1% for ViT-B-32. However, the minority-class performance remains weak, contrasting with findings in [4], where the degradation among rare classes was less severe. This difference suggests that CLIP's contrastive pretraining behaves differently on LSCIDMR-S, where high inter-class similarity especially between LabelLess and nearby cloud types causes feature entanglement that negatively impacts rare classes.

The Class-Balanced loss experiments produce more balanced performance across classes. ResNet-50 reaches 73.3% accuracy and ViT-B-32 reaches 74.8%, with macro-level metrics improving to 75.8% and 82.4%, respectively. These outcomes show a stable intermediate pattern between LA and CE, aligning with long-tailed learning insights from [29], which highlight CB loss

DOI: 10.31253/te.v9i2.4256

as an effective compromise between overall accuracy and class fairness. The reduced performance gap between LabelLess and the remaining classes particularly in ViT-B-32 demonstrates the loss function's suitability for handling highly skewed cloud distributions.

Reducing τ in Logit Adjustment leads to only partial improvements, increasing overall accuracy to 75.3% but decreasing macro-level performance, indicating that the imbalance effect is not fully resolved. Focal Loss yields moderate improvements at 81.4% accuracy but still underperforms compared to CE and CB, deviating from the expectations outlined in [30], where Focal Loss improved minority recognition in other imbalanced scenarios. LADE delivers the weakest performance among alternative losses at 55.7%, confirming that some imbalance-focused methods do not generalize well to meteorological cloud imagery.

Substantial performance gains emerge when the LabelLess class is removed. Under this setting, ViT-B-32 achieves 90.6% accuracy using LA and 91.2% using CB. The best performance in this subset is achieved by CB with a 128-dimensional bottleneck, reaching 93.6% accuracy and a macro F1 of 91.1%. Although these results approach previously reported accuracies of 93–97% [4], [5], [17], they remain slightly lower due to the parameter-efficient tuning strategy used in this work, whereas prior studies relied on fully supervised CNN architectures with larger parameter budgets.

Increasing the adapter bottleneck dimension provides additional insights. Expanding from 32 to 64 dimensions increases accuracy to 62.5%, while a 128-dimensional bottleneck yields 63.7%, illustrating diminishing returns. When the 128-dimensional adapter is paired with CB loss, accuracy rises to 78.5% with a strong macro F1 of 70.6%. With CE at 128 dimensions, accuracy increases further to 84.2%, nearly matching earlier studies (84.5%), although at the cost of reduced per-class fairness. Extending training to 15 epochs results in slight but consistent performance gains, increasing accuracy to 79.4% and macro F1 to 71.8%.

Overall, the results demonstrate that no single configuration optimizes every performance metric simultaneously. However, ViT-based models consistently outperform CNN-based models across nearly all experiments, particularly when paired with Class-Balanced loss and larger adapter bottlenecks. While earlier high-performing CNN models primarily emphasized maximizing total accuracy, this study shows that foundation-model-based ViTs provide more equitable performance across both frequent and rare cloud classes. Such balanced recognition is essential in meteorological applications, where rare but high-impact weather phenomena require reliable detection.

After evaluating all models, the best-performing model is deployed in a web interface capable of predicting weather systems from satellite cloud imagery. Figure 6 shows the landing page of the website, where the background image is an RGB satellite image from Himawari-9 provided by the Japan Meteorological Agency under a CC BY 4.0 International license. Users can click a button to upload an image and generate predictions. Figure 7 displays the prediction page showing the uploaded image along with its prediction probabilities. Figure 8 shows the page displayed when the information icon is clicked, providing an explanation of the predicted results. Figure 9 illustrates the pop-up that appears when a sample image is selected.

Figure 6. Landing Page

Figure 7. Prediction Page

DOI: 10.31253/te.v9i2.4256

Figure 8. Explanation Page

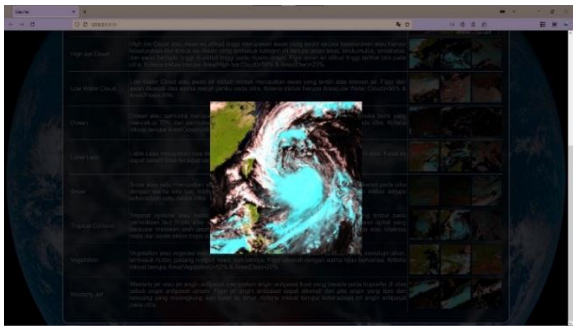


Figure 9. Pop-up Page

These findings have broader implications for the global scientific community and society at large. First, improving classification balance across cloud types is essential for supporting more reliable automated weather monitoring systems worldwide. As extreme weather events increase due to climate change, balanced model performance ensures that rare but impactful weather systems such as deep convection, typhoons, or frontal boundaries are not overlooked.

Second, the study demonstrates how foundation models can enhance meteorological AI applications globally, particularly in regions with limited forecasting resources. Lightweight parameter-efficient tuning allows the deployment of advanced models without requiring high-compute infrastructure, making satellite-based nowcasting more accessible to developing countries.

Third, the research highlights the importance of long-tailed learning for Earth observation more broadly. Many environmental datasets naturally exhibit skewed distributions, and the demonstrated effectiveness of Class-Balanced loss and ViT-based architectures can inform future work in fields such as climate monitoring, disaster detection, and ecological surveillance.

Finally, by integrating the best-performing model into a publicly accessible web application, this work supports global efforts toward democratizing access to meteorological intelligence and improving societal resilience to extreme weather conditions.

VII. CONCLUSION

This study successfully developed and evaluated CNN- and ViT-based models for long-tailed satellite cloud image classification using the LSCIDMR-S dataset. The experiments highlighted that severe class imbalance presents significant challenges to model performance, especially when using default configurations like Logit Adjustment. Through systematic optimization, including adjustments to loss functions and parameter-efficient tuning strategies, model stability and accuracy were significantly improved. Among the tested architectures, ViT-based models consistently outperformed CNN-based models, achieving superior overall and per-class performance, closely aligning with results from earlier cloud-classification studies.

The introduction of larger adapter bottleneck dimensions further enhanced ViT performance, underscoring the importance of increased representational capacity in long-tailed scenarios. Additionally, Class-Balanced loss emerged as the most effective learning objective, offering a balanced trade-off between overall accuracy and equitable performance across both rare and frequent classes. These findings demonstrate the potential of foundation-model–based ViT architectures, paired with appropriate loss functions and tuning strategies, to provide robust and well-balanced classification for satellite cloud imagery in highly imbalanced real-world settings.

VIII. ACKNOWLEDGEMENT

The authors would like to express our gratitude for the support provided by Faculty of Science and technology. The financial assistance from Research, Publication and Community Service Department Buddhi Dharma University is also greatly acknowledged. (Use acknowledgement if research uses UBD financial assistance)

DOI: 10.31253/te.v9i2.4256

REFERENCES

[1] T. B. Turrisi *et al.*, "Seasons, weather, and device-measured movement behaviors: a scoping review from 2006 to 2020," *Int. J. Behav. Nutr. Phys. Act.*, vol. 18, no. 1, p. 24, Feb. 2021, doi: 10.1186/s12966-021-01091-1.

[2] B. Clarke, F. Otto, R. Stuart-Smith, and L. Harrington, "Extreme weather impacts of climate change: an attribution perspective," *Environ. Res. Clim.*, vol. 1, no. 1, p. 012001, Sep. 2022, doi: 10.1088/2752-5295/ac6e7d.

[3] S. Fawzy, A. I. Osman, J. Doran, and D. W. Rooney, "Strategies for mitigation of climate change: a review," *Environ. Chem. Lett.*, vol. 18, no. 6, pp. 2069–2094, Nov. 2020, doi: 10.1007/s10311-020-01059-w.

[4] C. Bai, M. Zhang, J. Zhang, J. Zheng, and S. Chen, "LSCIDMR: Large-Scale Satellite Cloud Image Database for Meteorological Research," *IEEE Trans. Cybern.*, vol. 52, no. 11, pp. 12538–12550, Nov. 2022, doi: 10.1109/TCYB.2021.3080121.

[5] S. Shang, J. Zhang, X. Wang, X. Wang, Y. Li, and Y. Li, "Faster and Lighter Meteorological Satellite Image Classification by a Lightweight Channel-Dilation-Concatenation Net," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 16, pp. 2301–2317, 2023, doi: 10.1109/JSTARS.2023.3243915.

[6] E. Chuvieco, *Fundamentals of Satellite Remote Sensing*. CRC Press, 2020. doi: 10.1201/9780429506482.

[7] F. A. Diaz-Gonzalez, J. Vuelvas, C. A. Correa, V. E. Vallejo, and D. Patino, "Machine learning and remote sensing techniques applied to estimate soil indicators – Review," *Ecol. Indic.*, vol. 135, p. 108517, Feb. 2022, doi: 10.1016/j.ecolind.2021.108517.

[8] W. Han *et al.*, "A survey of machine learning and deep learning in remote sensing of geological environment: Challenges, advances, and opportunities," *ISPRS J. Photogramm. Remote Sens.*, vol. 202, pp. 87–113, Aug. 2023, doi: 10.1016/j.isprsjprs.2023.05.032.

[9] F. Li, T. Yigitcanlar, M. Nepal, K. Nguyen, and F. Dur, "Machine learning and remote sensing integration for leveraging urban sustainability: A review and framework," *Sustain. Cities Soc.*, vol. 96, p. 104653, Sep. 2023, doi: 10.1016/j.scs.2023.104653.

[10] M. Marjani, M. Mahdianpari, F. Mohammadimanesh, and E. W. Gill, "CVTNet: A Fusion of Convolutional Neural Networks and Vision Transformer for Wetland Mapping Using Sentinel-1 and Sentinel-2 Satellite Data," *Remote Sens.*, vol. 16, no. 13, p. 2427, Jul. 2024, doi: 10.3390/rs16132427.

[11] M. Segal-Rozenhaimer, A. Li, K. Das, and V. Chirayath, "Cloud detection algorithm for multi-modal satellite imagery using convolutional neural-networks (CNN)," *Remote Sens. Environ.*, vol. 237, p. 111446, Feb. 2020, doi: 10.1016/j.rse.2019.111446.

[12] A. Galdran, G. Carneiro, and M. A. G. Ballester, "Convolutional Nets Versus Vision Transformers for Diabetic Foot Ulcer Classification," Nov. 2021, doi: 10.48550/arXiv.2111.06894.

[13] Z. Zhang, M. Lu, S. Ji, H. Yu, and C. Nie, "Rich CNN Features for Water-Body Segmentation from Very High Resolution Aerial and Satellite Imagery," *Remote Sens.*, vol. 13, no. 10, p. 1912, May 2021, doi: 10.3390/rs13101912.

[14] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," Jun. 2021, [Online]. Available: http://arxiv.org/abs/2010.11929

[15] M. Kaselimi, A. Voulodimos, I. Daskalopoulos, N. Doulamis, and A. Doulamis, "A Vision Transformer Model for Convolution-Free Multilabel Classification of Satellite Imagery in Deforestation Monitoring," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 34, no. 7, pp. 3299–3307, Jul. 2023, doi: 10.1109/TNNLS.2022.3144791.

[16] R. Rad, "Vision Transformer for Multispectral Satellite Imagery: Advancing Landcover Classification," in *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, IEEE, Jan. 2024, pp. 8161–8168. doi: 10.1109/WACV57701.2024.00799.

DOI: 10.31253/te.v9i2.4256

[17] R. Yousaf *et al.*, "Satellite Imagery-Based Cloud Classification Using Deep Learning," *Remote Sens.*, vol. 15, no. 23, p. 5597, Dec. 2023, doi: 10.3390/rs15235597.
[18] J.-X. Shi, T. Wei, Z. Zhou, J.-J. Shao, X.-Y. Han, and Y.-F. Li, "Long-Tail Learning with Foundation Model: Heavy Fine-Tuning Hurts," Jun. 2024, [Online]. Available: http://arxiv.org/abs/2309.10019
[19] F. Liu *et al.*, "RemoteCLIP: A Vision Language Foundation Model for Remote Sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–16, 2024, doi: 10.1109/TGRS.2024.3390838.
[20] R. Reedha, E. Dericquebourg, R. Canals, and A. Hafiane, "Transformer Neural Network for Weed and Crop Classification of High Resolution UAV Images," *Remote Sens.*, vol. 14, no. 3, pp. 1–20, 2022, doi: 10.3390/rs14030592.
[21] Z. Zhang, T. Zhao, Y. Guo, and J. Yin, "RS5M and GeoRSCLIP: A Large-Scale Vision-Language Dataset and a Large Vision-Language Model for Remote Sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–23, 2024, doi: 10.1109/TGRS.2024.3449154.
[22] S. Kim, L. Chen, and J. Kim, "Intrusion Prediction using LSTM and GRU with UNSW-NB15," in *2021 Computing, Communications and IoT Applications (ComComAp)*, 2021, pp. 101–106. doi: 10.1109/ComComAp53641.2021.9652926.
[23] M. Vlaminck, R. Heidbuchel, W. Philips, and H. Luong, "Region-Based CNN for Anomaly Detection in PV Power Plants Using Aerial Imagery," *Sensors*, vol. 22, no. 3, pp. 1–18, 2022, doi: 10.3390/s22031244.
[24] Z. Wang, J. Zhao, R. Zhang, Z. Li, Q. Lin, and X. Wang, "Uatnet: U-shape attention-based transformer net for meteorological satellite cloud recognition," *Remote Sens.*, vol. 14, no. 1, 2022, doi: 10.3390/rs14010104.
[25] Y. Tong, W. Lu, Y. Yu, and Y. Shen, "Application of machine learning in ophthalmic imaging modalities," *Eye Vis.*, vol. 7, no. 1, pp. 1–15, 2020, doi: 10.1186/s40662-020-00183-6.
[26] V. R. Joseph, "Optimal ratio for data splitting," *Stat. Anal. Data Min. ASA Data Sci. J.*, vol. 15, no. 4, pp. 531–538, Aug. 2022, doi: 10.1002/sam.11583.
[27] A. E. Berndt, "Sampling Methods," *J. Hum. Lact.*, vol. 36, no. 2, pp. 224–226, May 2020, doi: 10.1177/0890334420906850.
[28] A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," Feb. 2021, [Online]. Available: http://arxiv.org/abs/2103.00020
[29] A. K. Menon, S. Jayasumana, A. S. Rawat, H. Jain, A. Veit, and S. Kumar, "Long-tail learning via logit adjustment," Jul. 2021, [Online]. Available: http://arxiv.org/abs/2007.07314
[30] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," Feb. 2018, [Online]. Available: http://arxiv.org/abs/1708.02002

DOI: 10.31253/te.v9i2.4256

BIOGRAPHY

Revatta Manggala Nandivadhano, student at Universitas Buddhi Dharma, class of 2021, with a specialization in machine learning and data analytics. His academic interests include the development of machine learning models, AI-based data analysis, and the application of intelligent computing algorithms to real-world problem-solving, particularly in information systems and data-driven decision-making. Throughout his studies, he has actively developed technical competencies through academic projects and applied research focused on improving the performance of digital systems and enhancing the efficiency of technology-based processes.

Aditiya Hermawan, full-time lecturer at the Faculty of Science and Technology and has been actively teaching since 2019. His academic expertise and research interests focus on machine learning, artificial intelligence, and the application of intelligent computing methods in data processing and system-based decision-making. In his academic activities, he is involved in teaching, research, and curriculum development that emphasizes the integration of theory and practice, particularly in the use of machine learning algorithms to address problems across various application domains. Additionally, he actively supervises students in research and final projects related to data analytics and intelligent systems, and contributes to the advancement of knowledge through scientific publications and other academic engagements.

Lidya Lunardi, graduate of the Informatics Engineering program at Universitas Buddhi Dharma, class of 2023. She is currently pursuing a Master's degree in Management at the same university. Her academic background and interests focus on machine learning, data analytics, and the utilization of artificial intelligence to support managerial decision-making and the development of data-driven systems. With a foundation in both informatics engineering and management, she is particularly interested in interdisciplinary approaches that integrate artificial intelligence technologies with business strategy and organizational management. Throughout her studies, she has actively developed her technical and analytical competencies through academic activities, research, and applied projects related to digital transformation and technology-driven innovation.

DOI: 10.31253/te.v9i2.4256