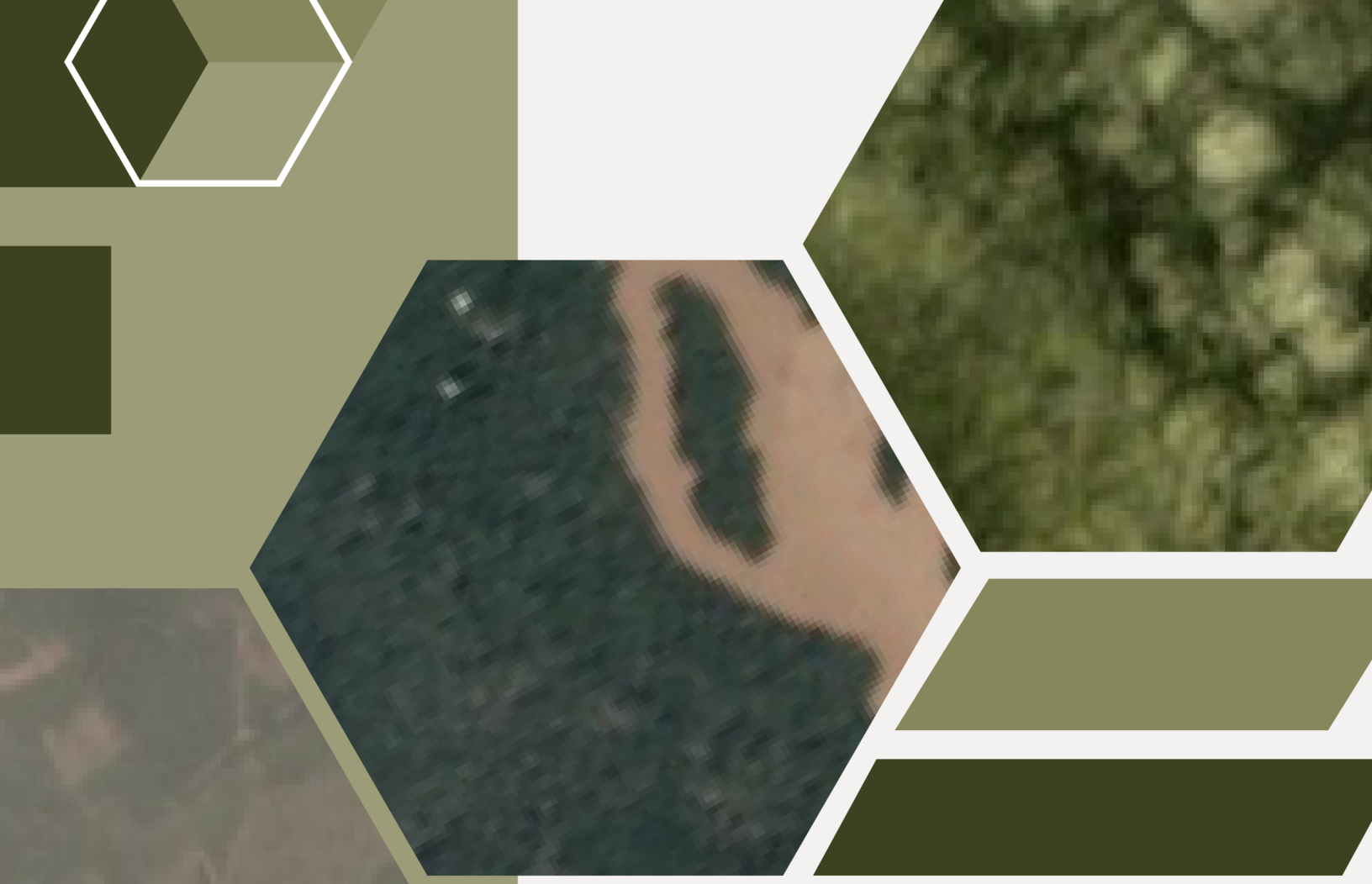


FINAL REPORT

Deep Learning for Amazon Rainforest Monitoring

Written By

Daianne Starr

Can a deep learning model identify and categorize environmental changes in the Amazon rainforest using multi-spectral satellite imagery?

PROBLEM STATEMENT

The Amazon rainforest is facing unprecedented levels of deforestation and degradation. The ability to quickly and accurately monitor changes in the rainforest is critical for conservation efforts and policy implementation. The problem is how to effectively process and analyze high-volume multi-spectral satellite imagery to detect and classify various environmental phenomena affecting the Amazon. This project sits at the intersection of environmental science and technology, leveraging the latest advancements in deep learning to provide actionable insights into the health of the Amazon rainforest

OBJECTIVES

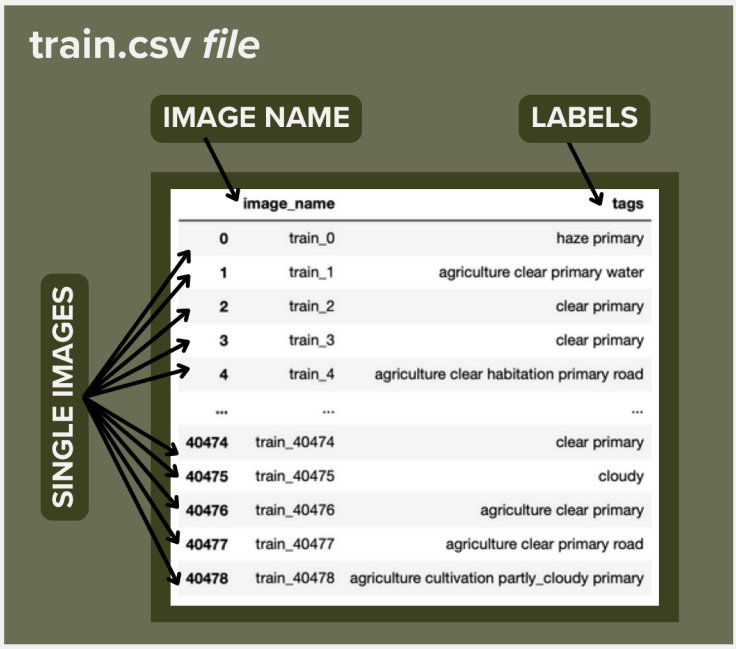
1. Develop a deep learning model that processes multi-spectral satellite imagery to accurately identify and classify various environmental phenomena in the Amazon rainforest, aiding in conservation and policy-making efforts.
2. Implement data preprocessing and augmentation techniques to enhance the model's ability to handle multi-label classification, addressing challenges such as class imbalance and the complex nature of environmental categorization.

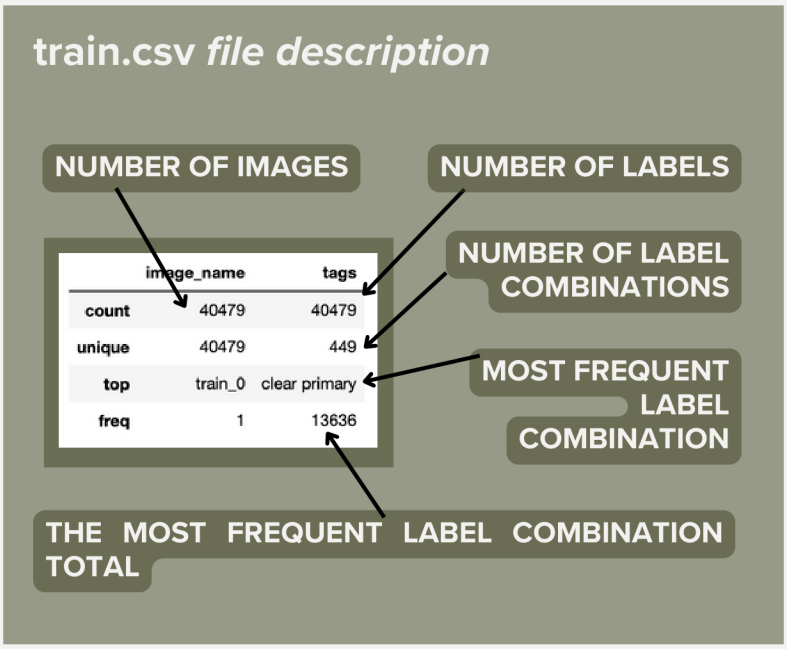
STAKEHOLDERS

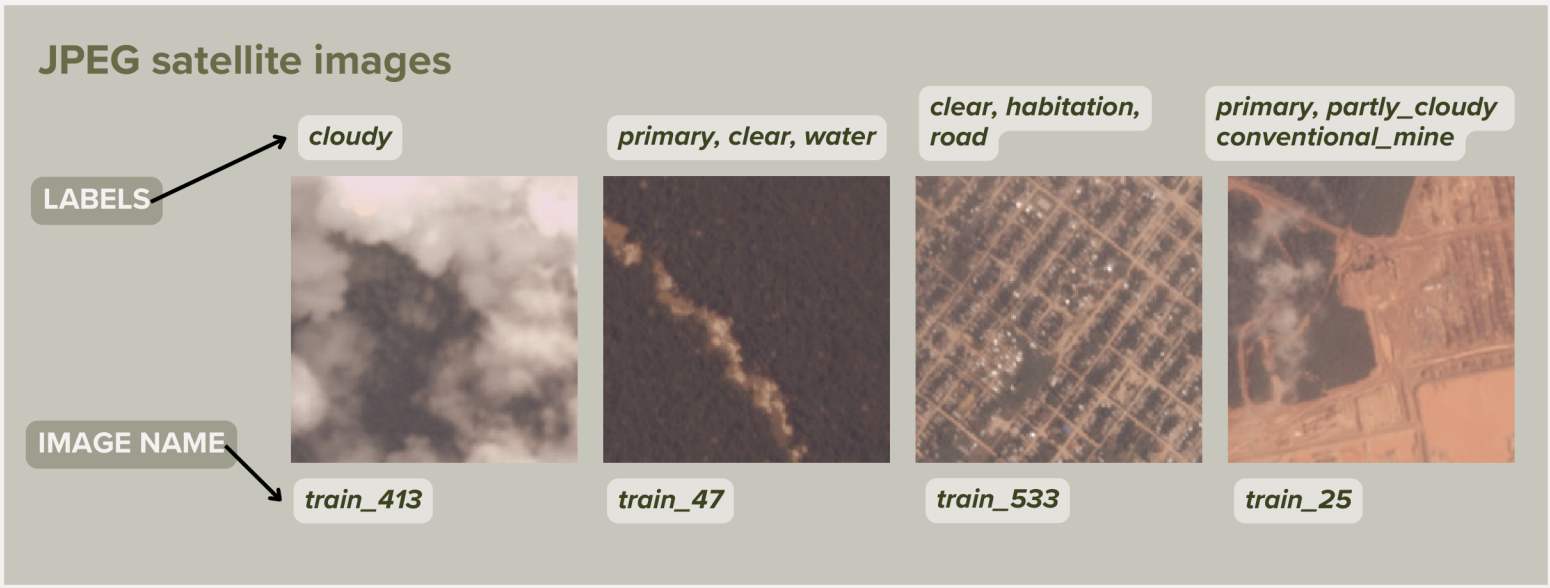
- Environmental NGOs and conservationists engaged in preserving the Amazon rainforest.
- Policymakers and governmental agencies responsible for managing and protecting the Amazon rainforest.
- Communities interested in engaging for the protection of the rainforest and addressing global climate change concerns.

DATASET

- The dataset for this project is sourced from Kaggle's 2017 Planet: Understanding the Amazon from Space competition. It features imagery from Planet's Flock 2 satellites, captured in both sun-synchronous and ISS orbits over the Amazon basin between 2016 and 2017. The images, with a ground-sample distance of 3.7m and an orthorectified pixel size of 3m, exclude geotiff details like chip footprint and ground control points.
- Covering regions from Brazil, Peru, Uruguay, Colombia, Venezuela, Guyana, Bolivia, and Ecuador, the dataset comprises **40,479** JPEG images. These are categorized into 17 classes such as clear, cloudy, and partially cloudy, agriculture, roads, blooming, artisanal mining. The accompanying **train.csv** file lists image names alongside their respective labels, providing a structured framework for model training and validation. This csv file contains 40,479 rows, each corresponding a different JPEG satellite image; and two columns: image_name and tags, which refers to to each JEPG image and its respective labels. The







DATASET WRANGLING

In the data wrangling phase of my project, I initially loaded the **train.csv** file as a DataFrame to effectively manage and manipulate the dataset. This file, containing image names and their corresponding multi-label tags, was crucial for the classification task. To better handle these labels, I transformed them into single categories for analysis. I achieved this by splitting the multi-label tags into lists using the **.str.split()** method, creating a new column **tags_list** in the DataFrame as seen in the **Figure A**.

	agriculture	artisinal_mine	bare_ground	blooming	blow_down	clear	cloudy	conventional_mine	cultivation	habitation	haze	partly_cloudy	primary	road	selective_logging	slash_burn	water
count	40479.00	40479.00	40479.00	40479.00	40479.00	40479.00	40479.00	40479.00	40479.00	40479.00	40479.00	40479.00	40479.00	40479.00	40479.00	40479.00	40479.00
mean	0.30	0.01	0.02	0.01	0.00	0.70	0.05	0.00	0.11	0.09	0.07	0.18	0.93	0.20	0.01	0.01	0.18
std	0.46	0.09	0.14	0.09	0.05	0.46	0.22	0.05	0.31	0.29	0.25	0.38	0.26	0.40	0.09	0.07	0.39
min	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25%	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
50%	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
75%	1.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
max	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Figure A: One-Hot Encoded DataFrame of Environmental Labels. This table represents the one-hot encoded transformation of the multi-label tags from the Amazon rainforest dataset. Each column corresponds to a unique environmental category, such as 'agriculture', 'clear', 'water', etc., with 40,479 entries. The table provides statistical insights including count, mean, standard deviation, and percentiles (min, 25%, 50%, 75%, max) for each category, reflecting the prevalence and distribution of these environmental features in the dataset.

Subsequently, I utilized the **MultiLabelBinarizer** from scikit-learn to convert these lists of tags into a one-hot encoded format. This transformation was pivotal in converting the multi-label problem into a binary format suitable for machine learning algorithms. The resulting one-hot encoded tags were then formed into a new DataFrame, allowing me to analyze and visualize the data more effectively. Lastly, I computed a correlation matrix from this DataFrame, which provided valuable insights into the relationships between different environmental labels in the dataset. This step was instrumental in understanding the interdependencies and patterns within the labels, guiding the further development of the deep learning model.

EXPLORATORY DATA ANALYSIS

HISTOGRAM

The initial stage of the EDA involved examining the frequency of occurrences for each label in the one-hot encoded dataset. The resulting histogram, shown in **Figure B**, is distinctly skewed with a pronounced long tail to the right. This tail indicates that a few labels occur more frequently than others, thus acting as outliers in this distribution.

BAR PLOTS

Subsequent analysis using bar plots of the same dataset, as presented in **Figures B-C**, further discriminates the frequency of the labels. Most of the label combinations occur less often across the images (**Figure B**). When the labels are separate (**Figure C**), the most common one is 'primary', representing the natural canopy of the Amazon forest, and it appears in 37,513 images. It is followed by 'clear', 'agriculture', 'road', and 'water', likely signifying the most prevalent features or conditions within the images.

IMBALANCE

Both the histogram and bar plots reveal an imbalanced distribution of labels in the dataset, a factor that could introduce bias in the model's development. The disproportionate quantity of 'primary' and 'clear' labels necessitates adjusting the model's weights to address the different frequencies of the classes, ensuring a balanced approach in the model's learning and prediction accuracy.

Figure B: Histogram depicting the frequency of occurrences for each unique label combination.

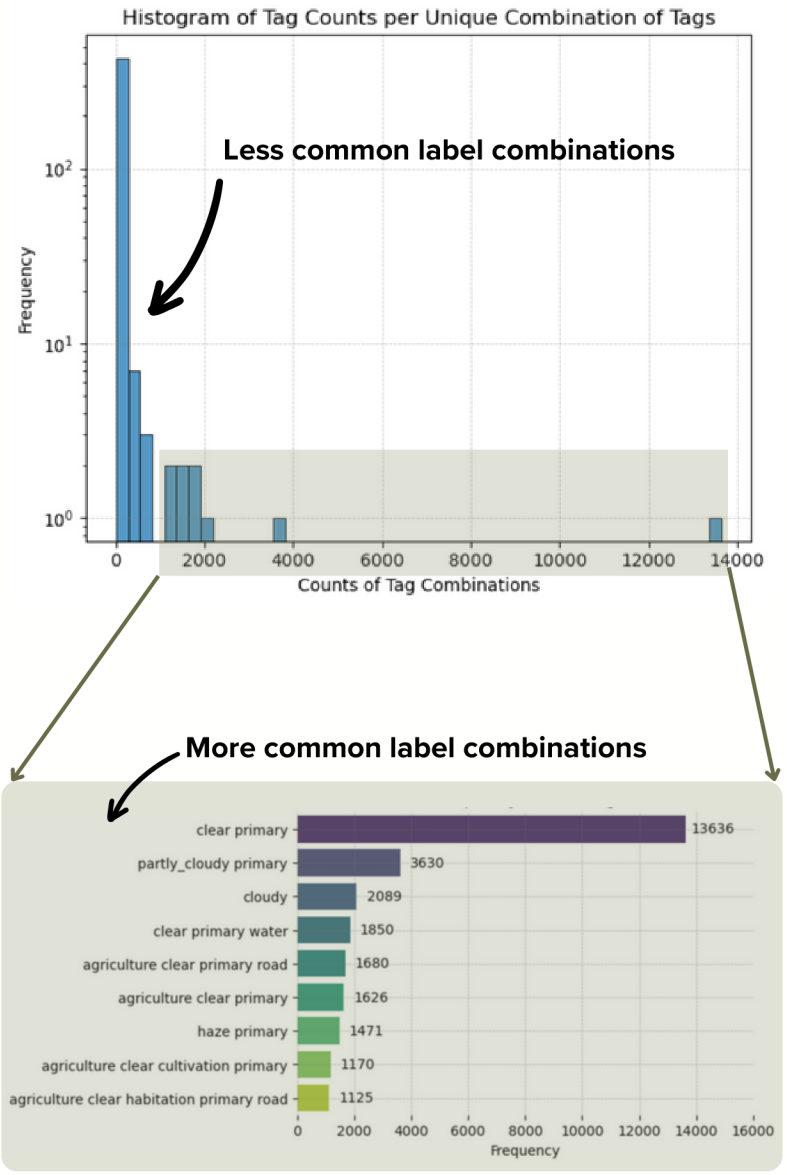


Figure C: Frequency distribution plot for each label combination above 1,000 counts derived from the satellite images.

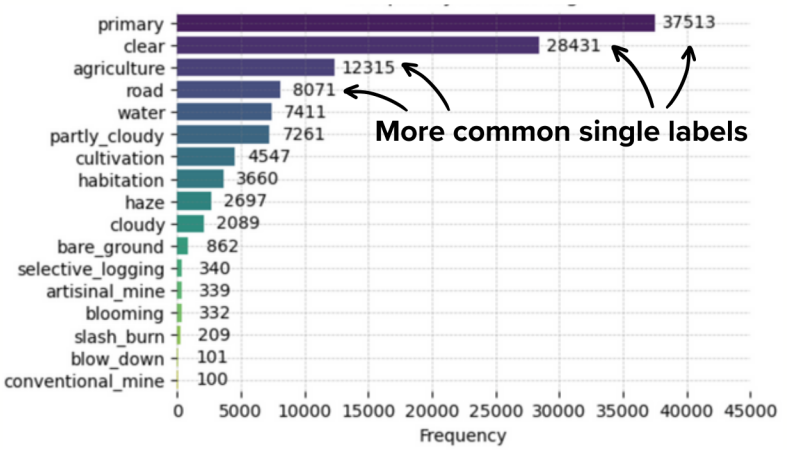
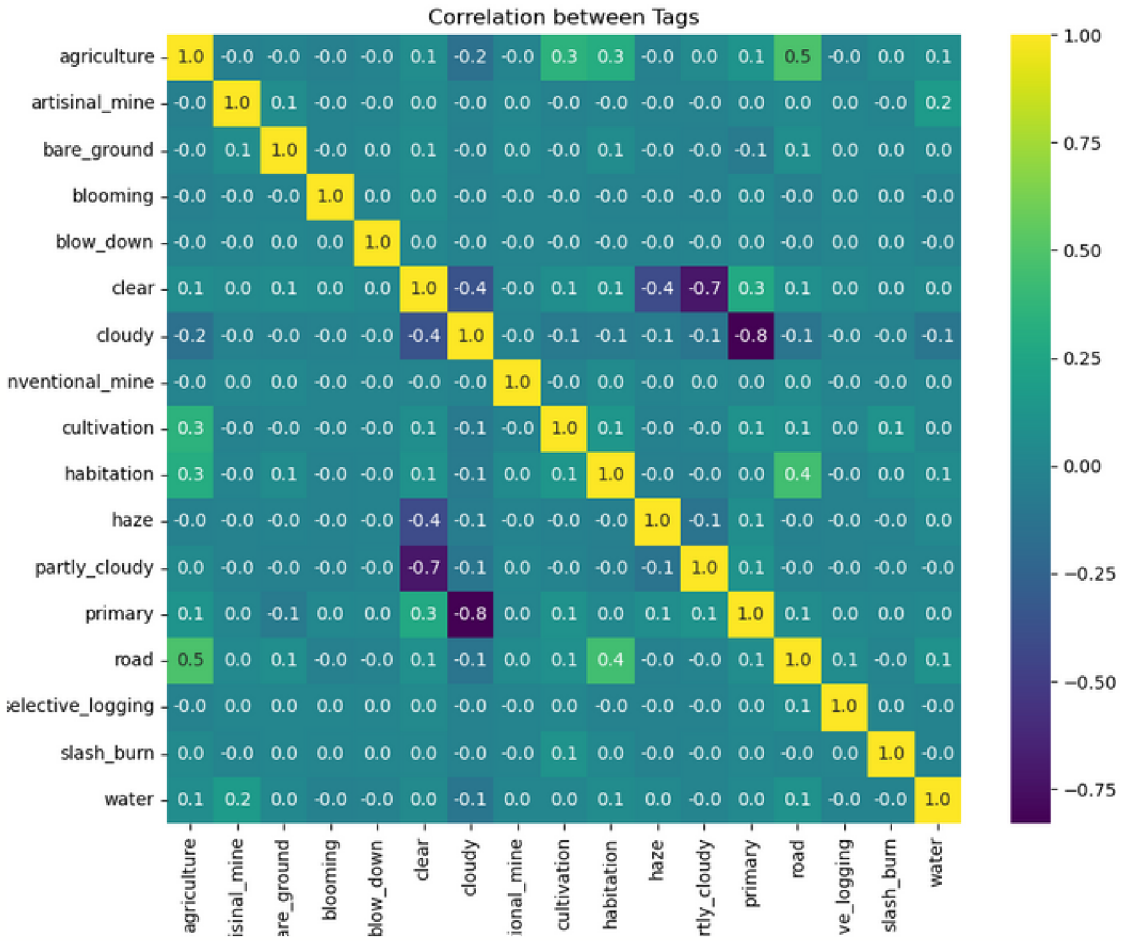


Figure D: Frequency distribution plot for each label derived from the satellite images.

Figure E: The heatmap reveals all the correlations among the labels occurring together in a same satellite image.



HEATMAPS

The general heatmap provides a visual representation of the Pearson correlation coefficients between the labels extracted from the satellite images as seen in **Figure C**. It reveals strong negative correlations between 'cloudy' and 'primary', as well as 'partially_cloudy' and 'clear', and between 'clear' and 'partially_cloudy'. Additionally, there are moderate positive correlations observed between 'agriculture' and 'road', 'habitation' and 'road', and between 'agriculture', 'habitation', and 'cultivation'. This insight into the relationships between different environmental conditions is crucial for understanding the complex dynamics within the satellite imagery data.

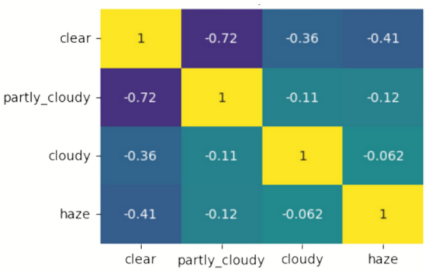
The associations between these labels are logical and coherent. For instance, the presence of the 'primary' label, indicating a clearly identifiable canopy, suggests that the conditions are not cloudy; cloud cover would otherwise hinder the identification of this feature. Similarly, in the context of land cover, the co-occurrence of 'settlements' and 'agricultural' regions with 'roads' is natural, as roads are essential for transportation in these areas. This inherent relationship between land use and accessibility is accurately reflected in the label correlations.

CO-OCCURENCE MATRICES

The EDA was refined by categorizing the single labels into specific groups. They included *weather*, *land_cover*, *human_activity*, and *natural_phenomenon* (**Figure F**). To ensure accuracy, the correlation matrix was recalculated using one-hot encoded data. Subsequently, a dictionary was created to store the correlation matrices filtered for each category.

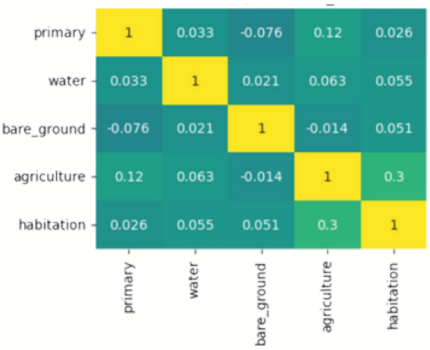
`. These category-specific correlation matrices were visualized using heatmaps. Each heatmap was annotated and color-coded to represent correlations ranging from -1 to 1, providing a clear and distinct representation of relationships within each category. Based on the co-occurrence matrix for each category:

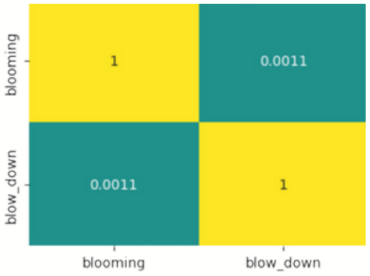
Figure F: Co-occurence matrices of the four different label groups: Weather, Land Cover, Activity and Phenomenon



Weather Labels: they describe the atmospheric conditions captured in the images. A strong negative correlation exists between *clear* and *partially_cloudy*, as well as between *haze* and *clear*. This suggests that these tags do not typically appear together in the images. Furthermore, it appears that each image is associated with a single weather label.

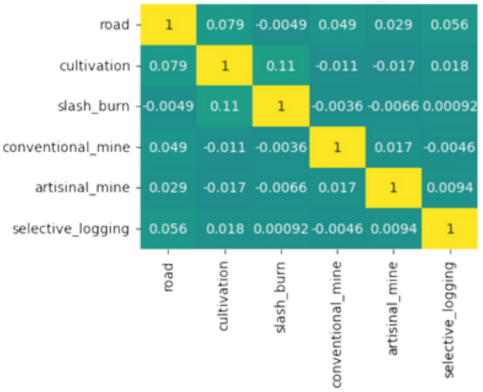
Land Cover Labels encompass tags related to the type of land cover, including water bodies and vegetation, as well as human-altered landscapes like agricultural areas and habitations. There is no clear linear relationship in the occurrence of these tags. The highest correlation coefficient observed is 0.3, which is between *agriculture* and *habitation*. However, this correlation is not particularly strong, suggesting that only a few images depict settled farmlands.

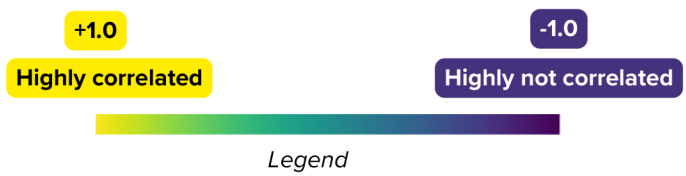




Natural Phenomenon Labels represent natural phenomena. These tags exhibit the lowest correlation values among all the categories, suggesting no discernible linear relationship in their occurrence. This lack of correlation is expected due to the random nature of such phenomena.

Human Activity Labels are associated with various human activities and their impact on the landscape, such as roads, agricultural cultivation, and slash-and-burn practices. The correlation values are close to zero, indicating no significant linear relationship in the occurrence of these tags.





AVERAGE RGB

In this phase of the EDA, the task was to extract and analyze color information from a dataset of images. Each image was separated into its Red, Green, and Blue (RGB) color channels, facilitating the individual visualization and examination of each channel to enhance the understanding of the images' color composition. A comprehensive dictionary linking each label with its average color values was created to achieve this goal. The horizontal bar chart of the average color distribution and pixel intensity (**Figure G**), contrasts the differences across labels. These RGB values, ranging from 0 (no intensity) to 255 (maximum intensity), offer insights into the general color characteristics of each labeled category in the image dataset.

Labels with higher average RGB values across all channels, such as cloudy, haze, artisinal_mine, and conventional_mine, suggest generally brighter images. This brightness could be indicative of lighter or more washed-out images, often associated with specific atmospheric conditions. Conversely, labels like primary, cultivation, slash_burn, and clear, characterized by lower RGB values, indicate darker images. This darkness could be attributed to features like dense foliage or canopy cover, typical in forest images. Special cases like *selective_logging*, *blow_down*, and *blooming* have the lowest RGB values, reflecting the darker nature of these images due to specific features of their categories.

Moreover, balanced RGB values in labels such as *partly_cloudy*, *road*, *habitation*, *agriculture*, and *water* suggest a more varied color palette within these categories, with no strong color dominance in any particular channel. Relative intensity differences between channels for certain labels, like lower blue values in *artisinal_mine* and *conventional_mine*, signal specific color tones or characteristics unique to those labels. This nuanced understanding of color distribution aids in grasping the unique visual attributes of each category.

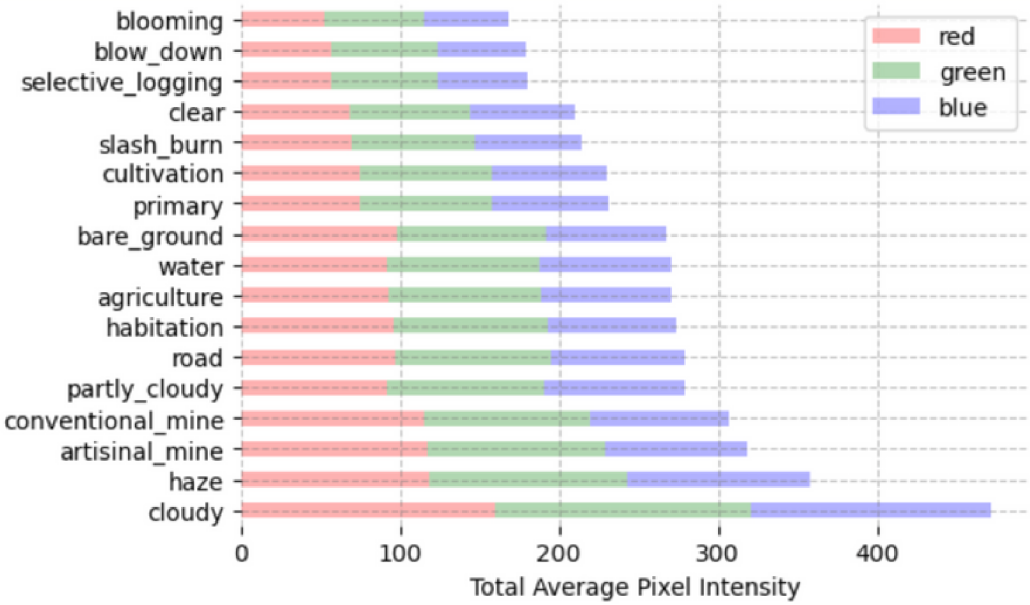


Figure G: Horizontal bar chart of average RGB color distribution and pixel intensity by label

AVERAGE IMAGES

In this analysis phase, the focus was on calculating and visually presenting the average images for each label in the dataset. The process involved iterating through each label and accumulating its associated images. For every label, this cumulative sum was divided by the total number of images, resulting in the average image for that label. These average images (**Figure H**) were then transformed into unsigned 8-bit integer arrays for clarity in visualization. Each image was annotated with its corresponding label and saved as a separate file. Each image in the grid represents the average color and texture characteristics of all images within its corresponding category.

For instance, *blooming* shows an average green hue typical for areas of trees, *clear* and *cloudy* could represent different weather conditions, with *clear* showing an average of clearer sky given by trees images and *cloudy* shows the average light gray tones of overcast conditions. Other labels display the average visuals that are not significantly distinct of each other. Each image in the grid is a data-driven visual summary, providing a quick, averaged glance at the typical visual features of each category—color distributions, possible textures, and overall brightness—that can be found in the full set of images under each label.

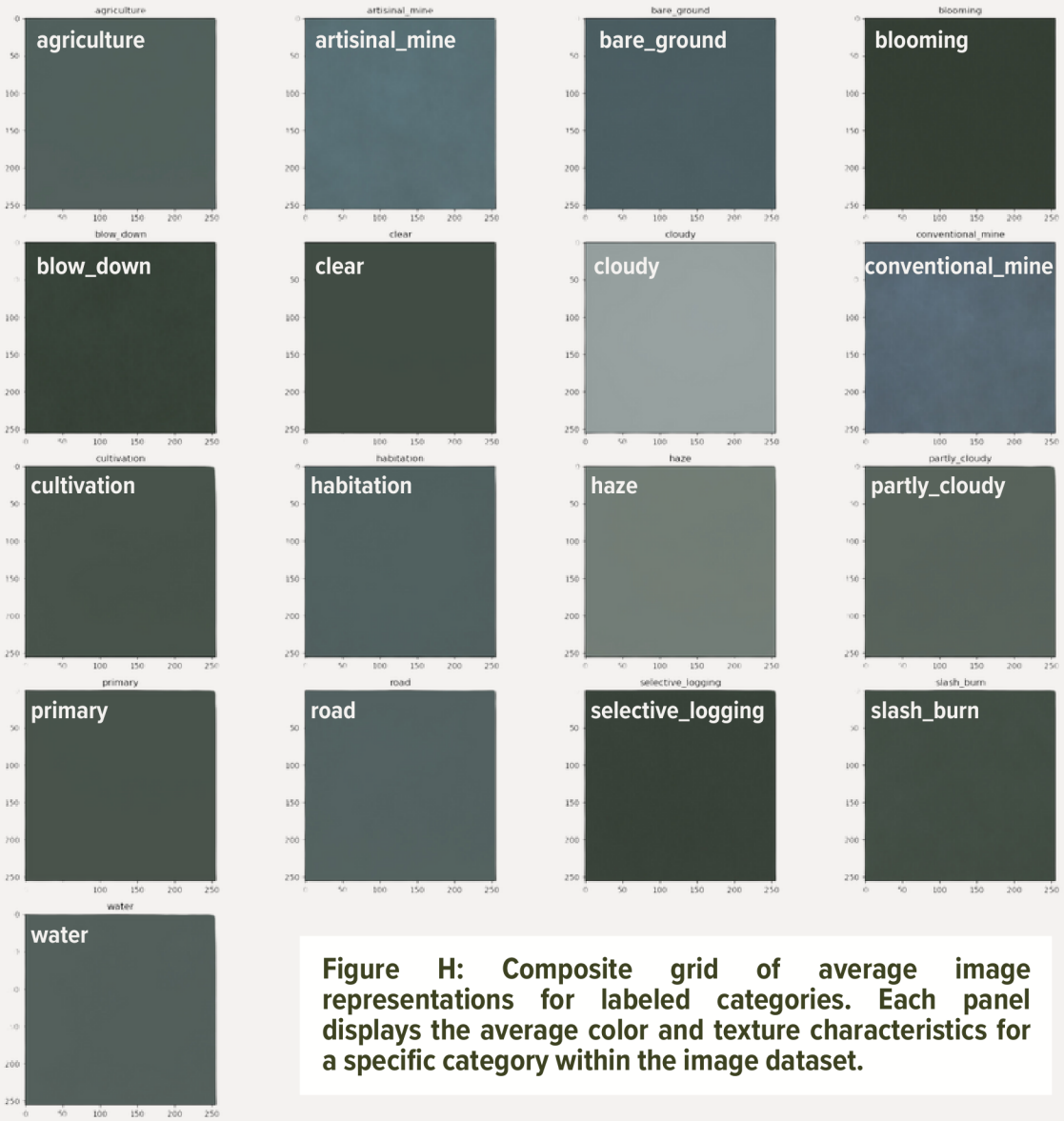


Figure H: Composite grid of average image representations for labeled categories. Each panel displays the average color and texture characteristics for a specific category within the image dataset.

CONTRAST BETWEEN AVERAGE IMAGES

This EDA step quantifies contrasts between average images, adopting Byeon's pixel difference methodology. By calculating and normalizing the absolute pixel value differences between image pairs, a visual comparison is achieved. The contrasts (**Figure I**) and illustrate variations in color and texture that are critical for image classification.

The provided visualizations succinctly capture the variance between satellite image labels, underscoring features essential for classification—like the textural and chromatic disparity between agricultural plots and water bodies. Contrasts between 'clear' and 'slash and burn' images highlight the environmental impact of such practices, while comparisons like 'bare ground' versus 'primary vegetation' or 'blooming' versus 'roads' detail the stark differences in natural versus human-altered landscapes.

The resulting contrast images are pivotal for discerning land cover types, revealing characteristic distinctions in texture and reflectance. This analysis enhances algorithmic classification by pinpointing distinct spectral signatures across the dataset, thereby advancing the understanding of satellite imagery for land categorization.

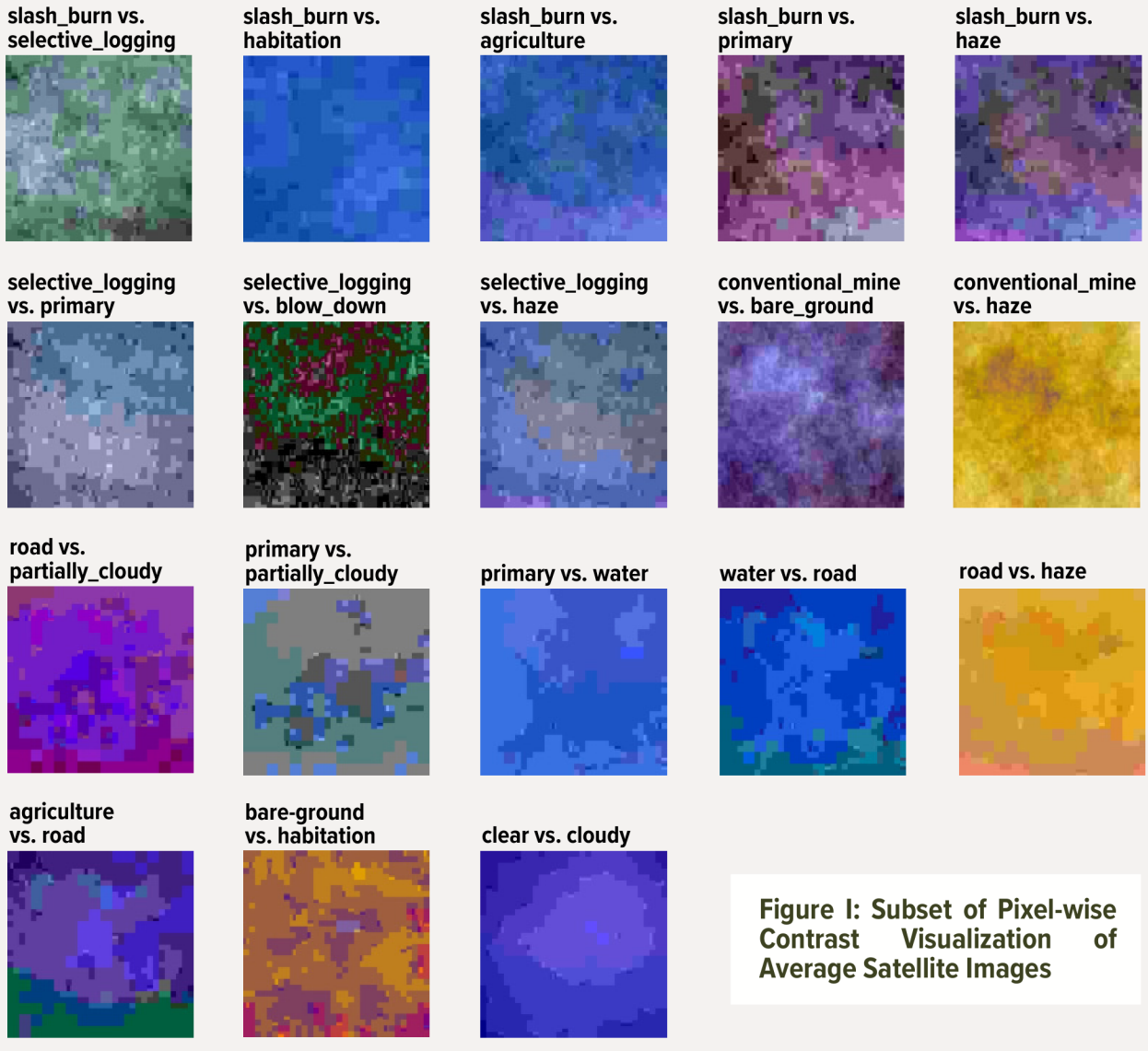


Figure I: Subset of Pixel-wise Contrast Visualization of Average Satellite Images

EIGENIMAGES

This analysis segment applies Principal Component Analysis (PCA) to extract Eigenimages from the satellite image dataset, is also a technique detailed in Byeon's work. The method select a random subset of images for each label, and processed it to grayscale to emphasize structural features. Utilizing the PCA algorithm, the code identifies the principal components that capture the most variance within these images (**Figure J**).

For example, in the context of satellite imagery, the first principal component (PC1) for an *agriculture* label capture the common patterns of agricultural fields, such as the egularity of crop rows or the specific texture of foliage. Similarly, the principal components for *cloudy* images reveal the density and patterns of cloud cover, while *habitation* capture the geometric regularity of human settlements. Variations within these components indicate different types of land use or environmental conditions.

In a comprehensive dataset, Eigenimages could help to identify features such as the ruggedness of *bare ground*, the irregular patches indicative of *slash and burn* practices, or the linear patterns of *roads*. By focusing on these Eigenimages, classifiers can be trained to recognize and categorize satellite images with higher accuracy, as they can learn from the essence of the data rather than noise or less informative variance.

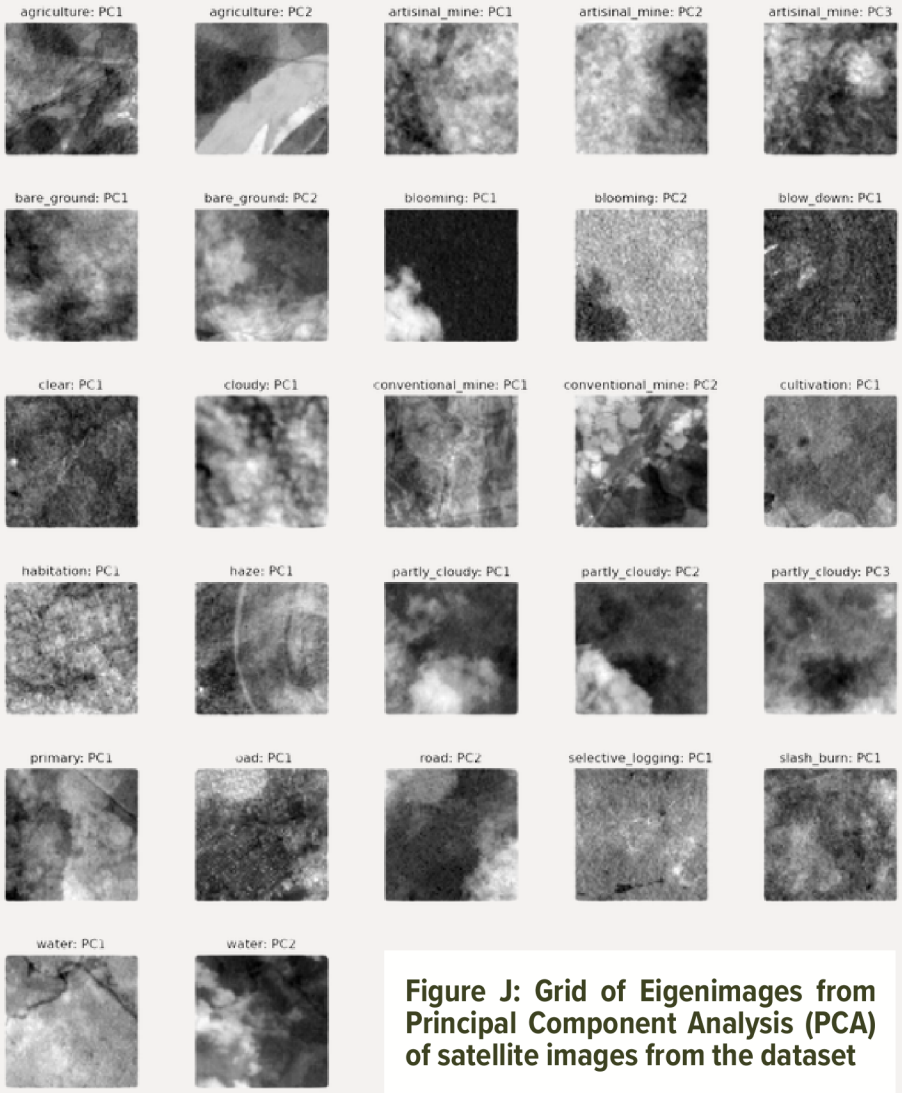


Figure J: Grid of Eigenimages from Principal Component Analysis (PCA) of satellite images from the dataset

PREPROCESSING

RESIZING

The preprocessing stage of the image classification project is multifaceted, focusing on resizing and dehazing the satellite images. Initially, the images are resized to 128x128 pixels from an original size of 256x256. This reduction is a trade-off to balance computational efficiency against the richness of image details, aiming to maintain sufficient quality for accurate model training while also expediting processing time.

ANTI-HAZE

Following resizing, an anti-haze filter is applied to clarify the images. The dehazing technique is inspired by He et al.'s method, which is particularly effective in improving the visibility of images obscured by haze. By calculating the minimum intensity across color channels, establishing a dark channel, and estimating atmospheric light, the algorithm attempts to recover the scene radiance. In doing so, it enhances the image by mitigating the effects of haze, which can obscure critical features and patterns necessary for accurate classification.

An additional refinement step is included in the form of Guided Image Filtering. This step enhances the dehazing process by leveraging the information contained in the original image to inform the refinement of the transmission map. The result is a clearer and more detailed image that retains the integrity of the original scene's structures and colors (**Figure K**).

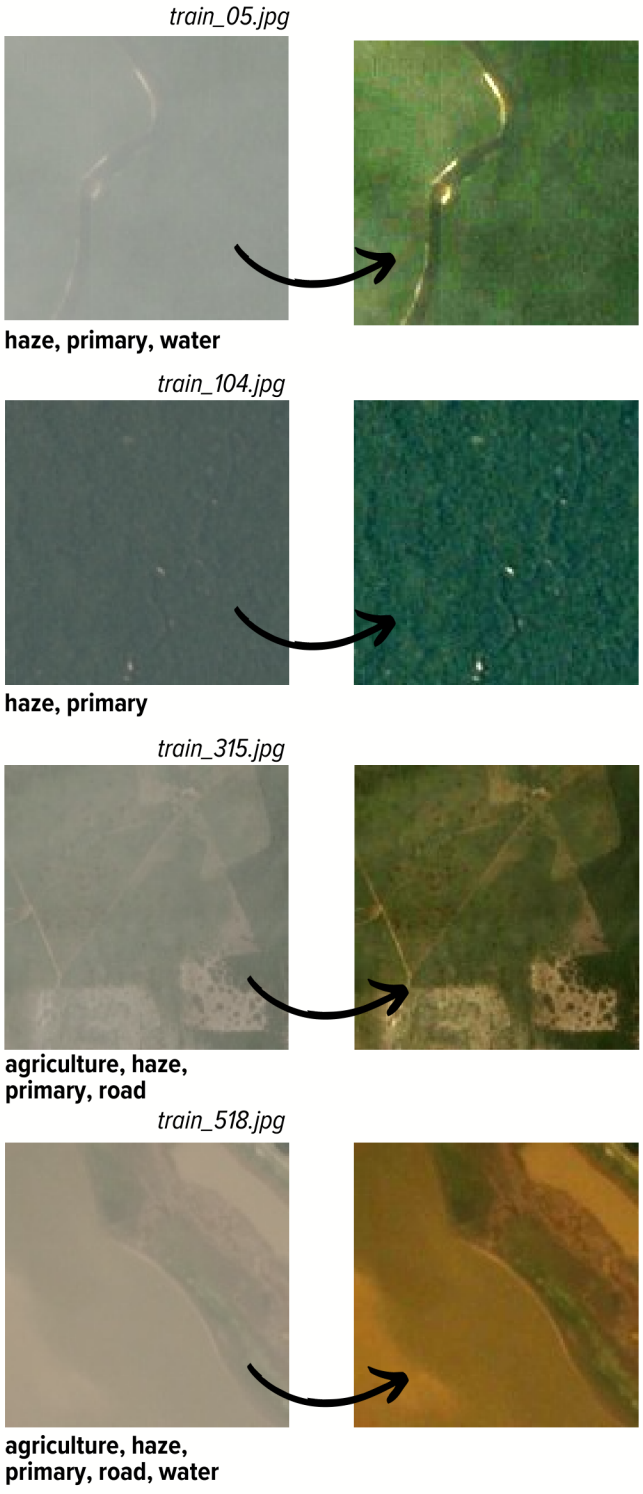


Figure K: Comparison of preprocessed satellite images before and after dehazing

MODELING

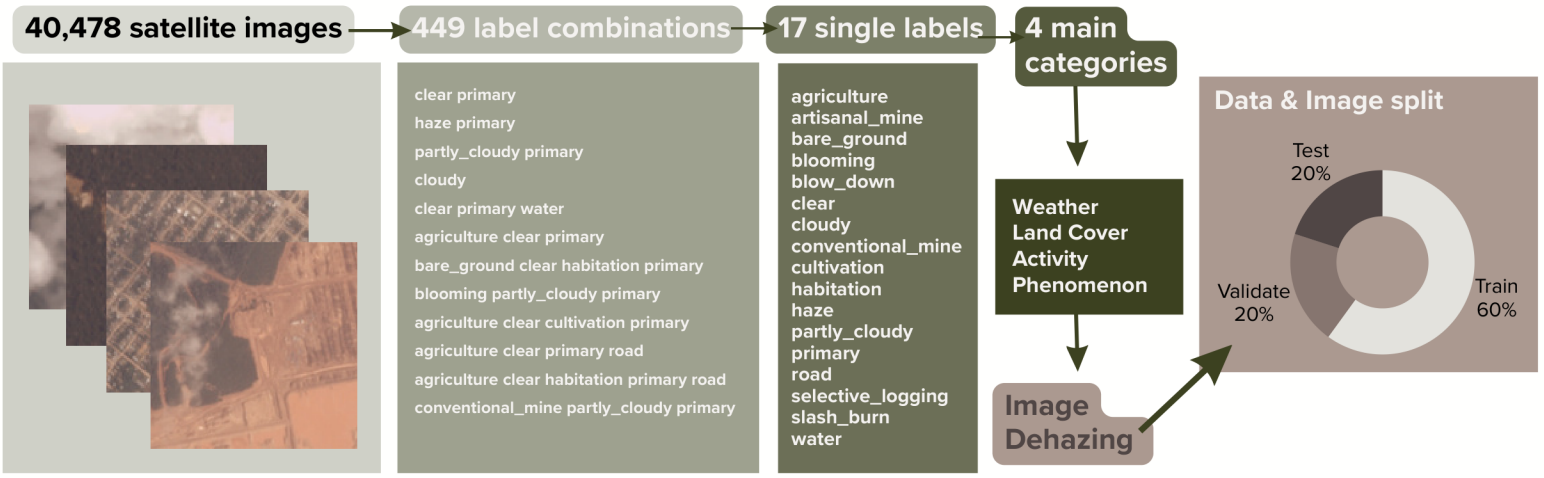
In order to construct a model capable to predict labels of satellite images from the Amazon Rainforest, a series of methodical steps were undertaken (**Figure L**). Firstly, grouped labels were split into arrays, followed by the implementation of one-hot encoding. This technique is pivotal in transforming categorical data into a format that is more conducive to machine learning algorithms. The resulting DataFrame meticulously collates the full paths to previously dehazed images along with their newly encoded tags.

Recognizing the potential for class imbalances within the dataset, a strategic step was taken to compute sample weights for each entry. This approach is instrumental in ensuring that rarer classes have a proportionately larger impact on the loss function during training. By doing so, it significantly mitigates the model's potential bias towards more frequently occurring classes.

Following this, a two-tiered data splitting strategy was employed. The dataset was initially divided into training and validation sets, adhering to a 60-40 split. The validation set was further segmented to create a test set, thereby establishing a tripartite division of the dataset into training, validation, and testing segments, accounting for 60%, 20%, and 20% of the dataset, respectively. The reindexing of these sets is a critical step, ensuring seamless data handling and referencing, which is indispensable during model training and evaluation.

To address memory constraints while loading images for classification during model training, validation, and testing, custom generators were developed. These generators play a pivotal role in efficiently feeding data in batches to the Keras model throughout the training and evaluation phases. They include capabilities for loading and preprocessing images, generating data for specific batch indexes, and shuffling these indexes at the end of each epoch, thereby ensuring a diverse range of training samples.

Figure L: In order to construct a model capable to predict labels of satellite images from the Amazon Rainforest, a series of methodical steps were undertaken



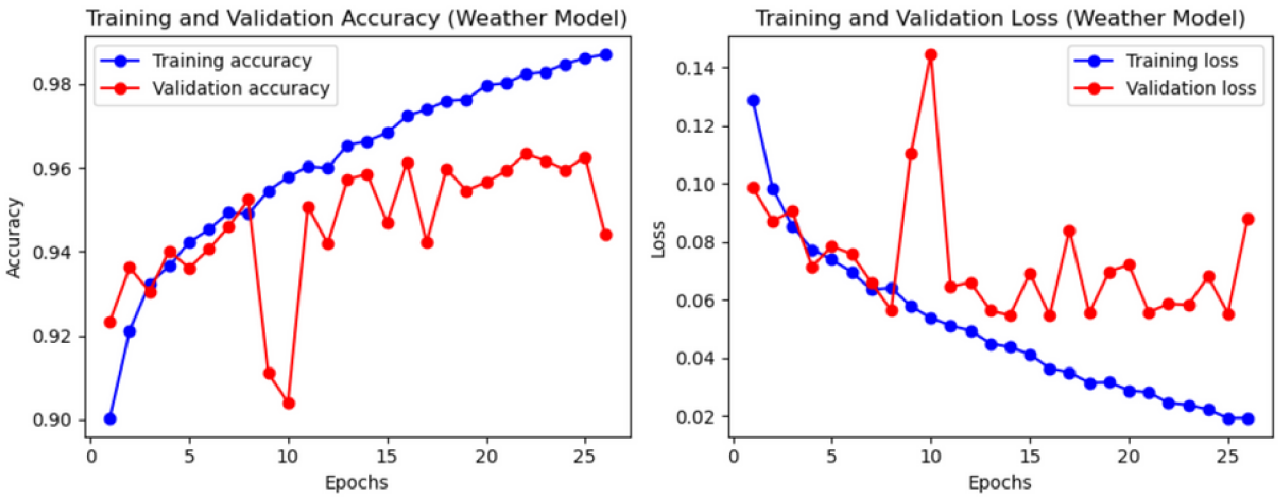
The model building function outlines the architecture of a convolutional neural network (CNN) using Keras' Sequential API. This architecture is characterized by three convolutional blocks, each comprising convolutional layers, batch normalization, ReLU activation, max-pooling, and dropout layers. These layers work collectively to extract features, reduce dimensionality, and prevent overfitting. The network concludes with dense layers, the final layer employing a sigmoid activation function, suitable for multi-label classification.

Four separate CNN models are then created and compiled, each tailored to predict a specific subset of classes (weather, land, activity, phenomenon) using binary cross-entropy as the loss function. This is appropriate for multi-label classification tasks. The models are trained using custom data generators for each category of labels, and early stopping and model checkpointing are utilized to enhance training efficiency.

The validation phase involves analyzing training and validation metrics from the history objects of the four models. Metrics include training loss, training accuracy, validation loss, and validation accuracy. These metrics are plotted to visually assess the models' performance, highlighting trends such as accuracy improvements and loss reductions. The fluctuating patterns in validation metrics across different models suggest areas for potential improvement, indicating the need for further model tuning.

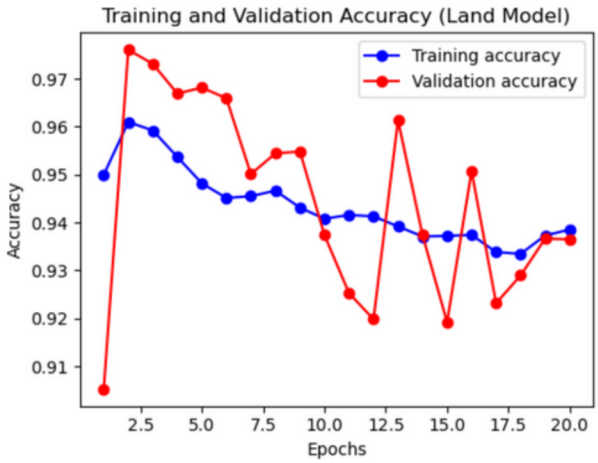
The use of subplots (**Figures M and N**) for each metric provides a comprehensive view of the models' performance, with color-coded visualizations aiding in the interpretation of the results. These plots offer valuable insights into the models' learning curves and generalization capabilities, guiding future refinements to enhance their predictive accuracy and reliability.

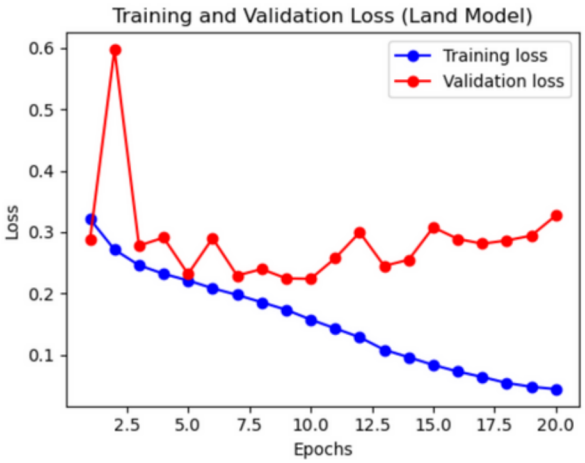
Figure M: Accuracy and validation for training and validation images and labels



Weather Model Metrics

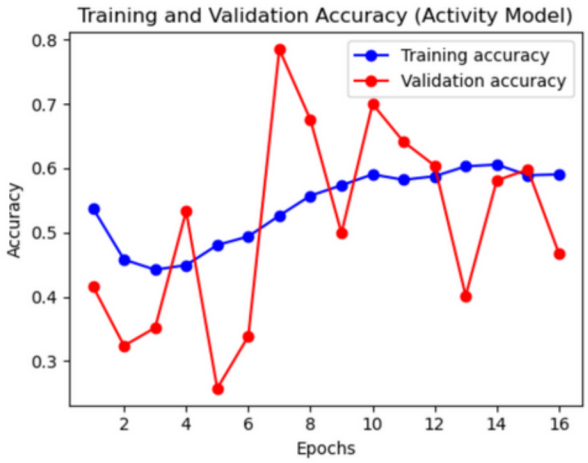
- The weather model's training accuracy rose notably, showing the model's growing proficiency in making correct predictions.
- Training loss decreased steadily, indicating it learned well from the training data.
- However, the validation loss fluctuated, suggesting variability in the model's generalization to new data. While validation accuracy generally increased, the observed fluctuations point to potential overfitting or the need for model tuning to ensure stable performance across various data sets.

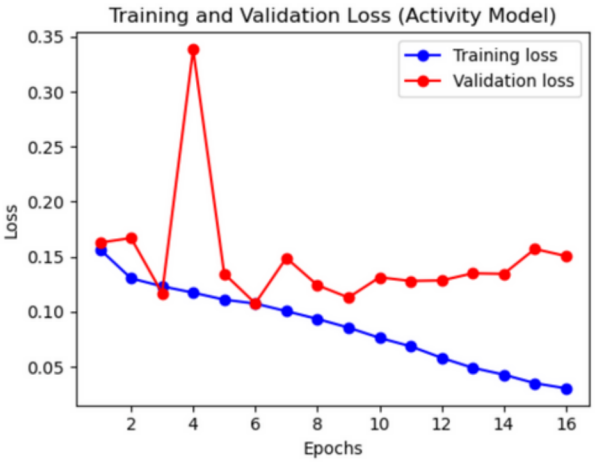




Land Model Metrics

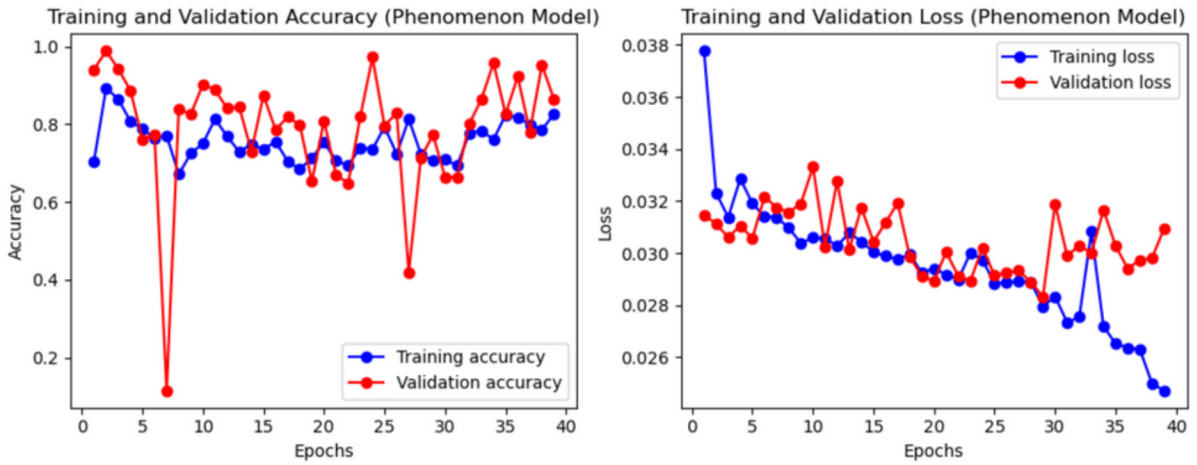
- The land model's high training accuracy suggests the model fits the training data well.
- The training loss illustrates a positive downward trend indicating that the model is effectively reducing error in its predictions as it learns.
- However, the validation loss presents a more complex picture with noticeable fluctuations and an upward trend towards the end of the training epochs, hinting at potential overfitting. The model might be too tailored to the training data, reducing its ability to generalize to new, unseen data.
- The validation accuracy mirrors this pattern of fluctuation but maintains high performance. Despite these peaks, the variability suggests that the model's ability to generalize may be inconsistent across different data subsets within the validation set.





Activity Model Metrics

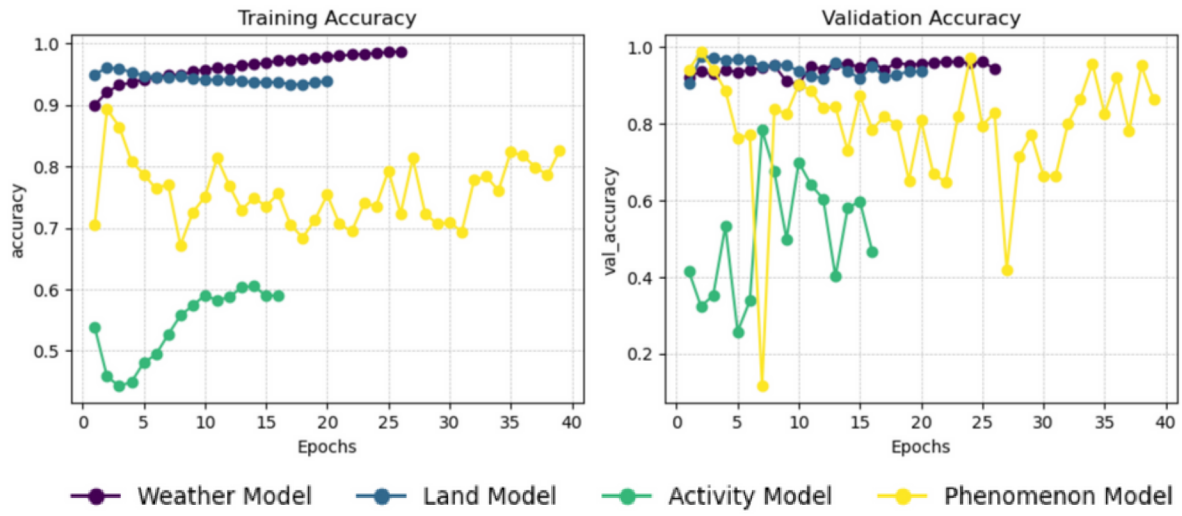
- The activity model's learning curve indicates substantial progress in model performance, with training loss decreasing notably, suggesting that the model is learning to classify the complex activity labels more effectively.
- Although the model's accuracy started relatively low, it shows a promising increase, pointing towards an improved ability to correctly predict activity labels as it processes more data.
- Validation results show variability but overall improvement, indicating better generalization to unseen data by the end of the training process. However, the fluctuations in validation loss hint at potential issues with model stability that could be addressed with further tuning.

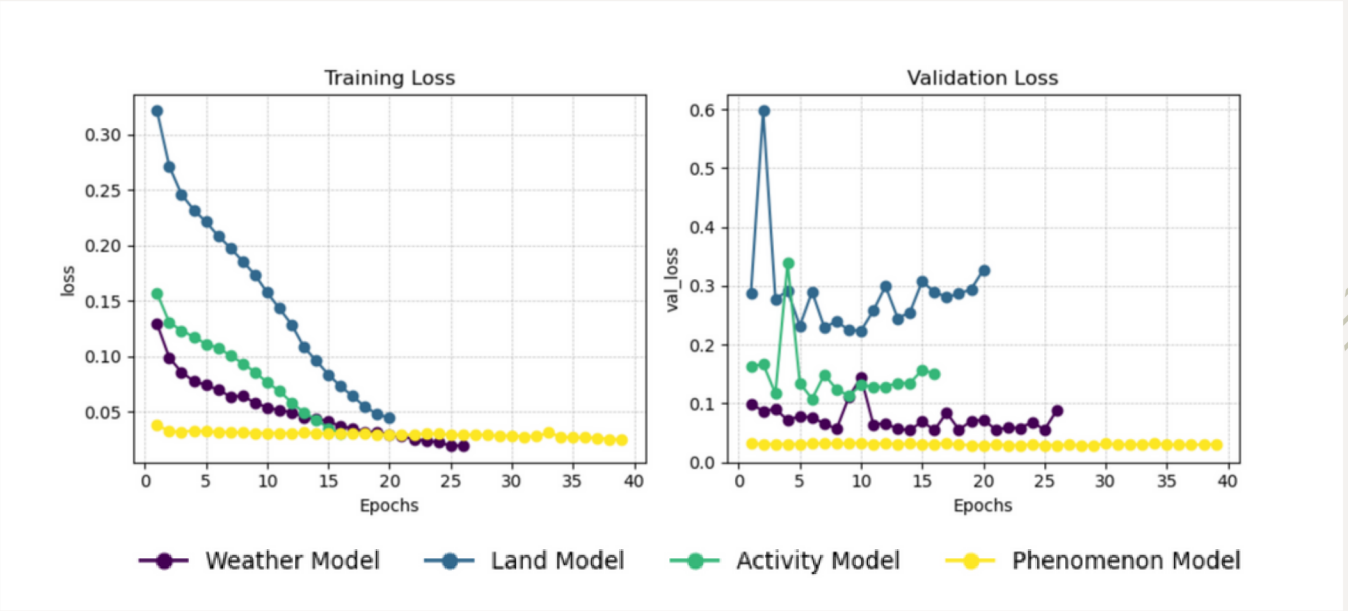


Phenomenon Model Metrics

- The phenomenon model displayed a commendable performance right from the outset, with a low training loss that decreased further, indicative of the model's proficiency in reducing error over time.
- Training accuracy experienced a rise, showcasing the model's capability to classify correctly during training sessions, despite minor fluctuations which are normal in the training process.
- In terms of validation, the loss remained minimal and showed a decrease, which is a strong sign of the model's ability to generalize to new, unseen data.
- The validation accuracy, while generally high, did exhibit some variability; which suggests that the model could effectively identify the correct phenomena in most cases.
- The variability in validation accuracy could point towards model sensitivity to specific data or potential overfitting, warranting further investigation and possible model refinement.

Figure N: Model accuracy and loss over epochs for the training and validation dataset. Figure continues in the next page.





Training Accuracy:

- The Activity Model starts at around 50% accuracy and shows a steady increase over the epochs, suggesting that the model is learning and improving its performance on the training set.
- The Phenomenon Model starts at around 80% accuracy and has a more variable progression but appears to reach a plateau, indicating the model might have reached its learning capacity on the training data.
- The Land Model line begins with high accuracy close to 90% and maintains a stable level throughout the training, which could suggest that this model was already well-tuned to the training data or perhaps is overfitting.

Validation Accuracy:

- The Activity Model shows high variability, with accuracy ranging from below 20% to above 60%, which suggests that the model might not be generalizing well to the validation set.
- The Phenomenon Model displays extreme variability, with accuracy swinging between below 20% to nearly 100%. This could indicate issues such as overfitting to the training data, a small validation set, or other problems that could cause the model to perform inconsistently on unseen data.
- The Weather Model's performance is relatively more stable than the activity and Phenomenon lines, but it also shows some decline, possibly indicating the beginning of overfitting as the epochs increase.

Training Loss:

- The four models show a decrease in training loss as the number of epochs increases. This is expected behavior, as a model's loss should generally decrease during training as it learns from the data.
- The Weather Model starts with the highest loss and shows a rapid decrease, flattening out as it approaches a loss close to zero, suggesting the model has learned well from the training data.

- The Weather and Activity models start at lower initial losses compared to the dark blue line and follow a similar rapid decline, flattening out as they approach a training loss of around 0.05 and 0.10 respectively.
- The Phenomenon Model starts with the lowest loss and decreases slightly over time, indicating that this configuration might have been close to optimal from the beginning or that there's less room for improvement given the model's complexity or the data's nature.

Validation Loss :

- The validation loss for the Land Model is relatively higher than its training loss, indicating a discrepancy between the model's performance on the training data versus unseen validation data, which could be a sign of overfitting.
- The Activity Model shows high variability in validation loss, with sharp peaks and troughs, suggesting that the model's performance on the validation set is unstable.
- The Weather Model has a relatively stable and low validation loss, suggesting that this model generalizes well to unseen data. However, there are a couple of spikes, which might be due to the variability of the validation data or the model's sensitivity to certain types of data.
- The Phenomenon Model maintains a low and stable validation loss throughout the training epochs, indicating that this model has good generalization and is the most consistent of the four.

In summary, these accuracy and loss graphs suggest that while the models are learning from the training data, there may be issues with how they perform on the validation data, which is crucial for evaluating the model's real-world applicability (**Figure O**). Strategies like adjusting model complexity, gathering more data, or employing regularization techniques might be needed to improve model generalization.

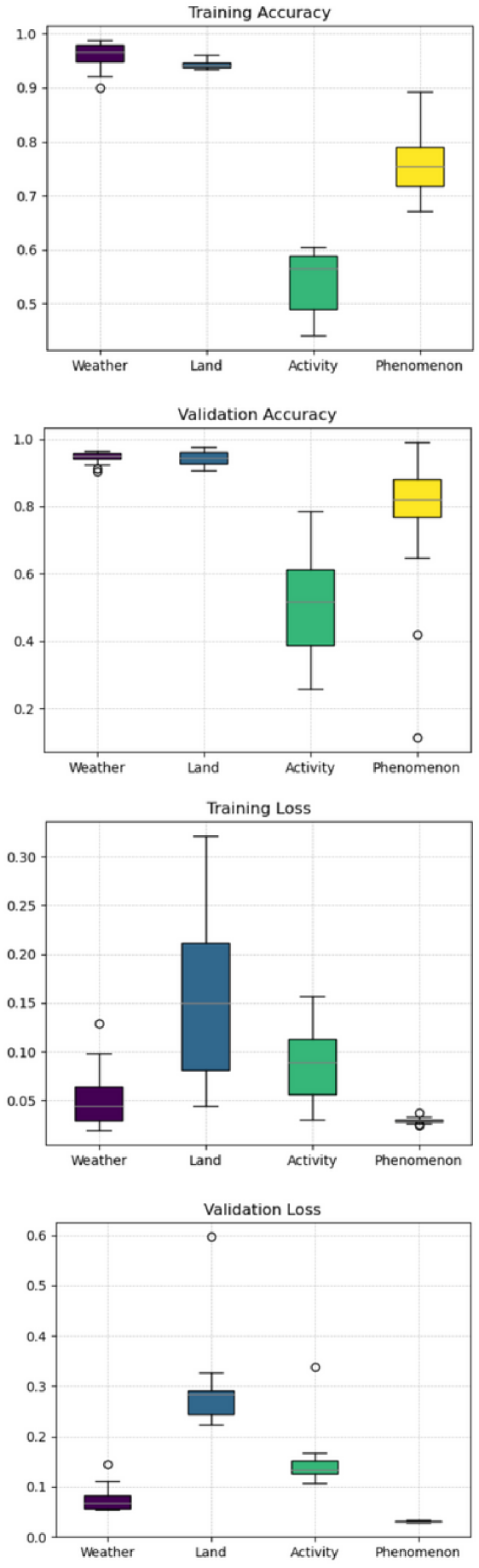


Figure O: Boxplots of training and validation accuracy and loss a by label category

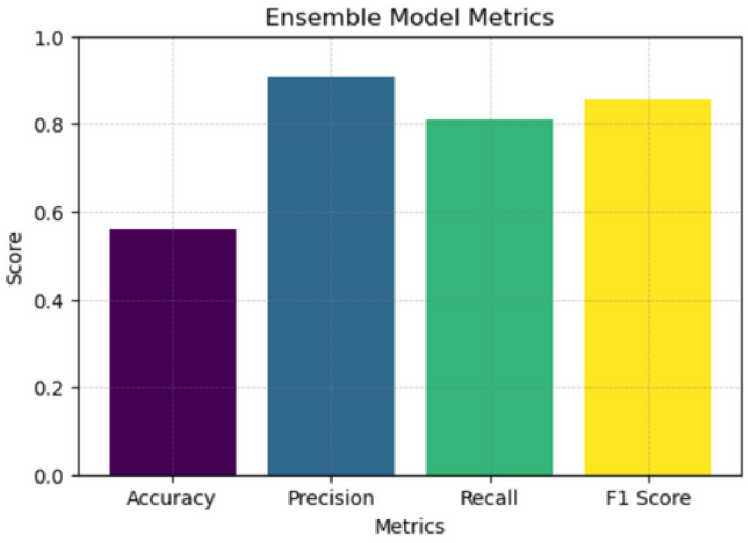


Figure O: Performance metrics of the ensemble model. This bar plot summarizes the ensemble model's accuracy, precision, recall, and F1 score on the test dataset.

In the testing phase, the model undergoes a rigorous evaluation to ascertain its predictive prowess on unseen data. The process begins with a standard image preprocessing routine: loading each image, resizing to a uniform dimension, converting the color space to RGB, and normalizing pixel values. Upon preprocessing, the model leverages its multiple specialized networks—each attuned to distinct tag categories like weather, land, activity, and phenomenon—to generate predictions for the test dataset.

These predictions are then meticulously aligned and aggregated, considering the unique indices assigned to each class, ensuring that predictions for different aspects of the data are appropriately combined. The ensemble approach, by integrating outputs from various networks, aims to harness the collective intelligence of the models, potentially improving overall performance.

Subsequently, the predictions are binarized using a threshold, distinguishing the presence or absence of each class. The binary predictions are juxtaposed against the actual labels of the test data, setting the stage for a comprehensive evaluation.

The binary predictions are juxtaposed against the actual labels of the test data, setting the stage for a comprehensive evaluation.

Metrics such as accuracy, precision, recall, and the F1 score are computed to provide a holistic view of the model's performance. These metrics serve as critical indicators of the model's effectiveness, with precision and recall offering insights into the model's true positive rate and its ability to minimize false negatives, respectively.

The bar plot shown in **Figure P** encapsulates the ensemble model's performance across these metrics. It reveals that while accuracy stands at 56.19%, precision soars to 90.79%, indicating a high rate of true positives among the predicted labels. The recall, at 81.19%, and the F1 score, at 85.72%, suggest a balanced trade-off between precision and recall, underpinning the model's robustness. The model's heightened precision and F1 score particularly underscore its capability to correctly identify and classify the varied and intricate patterns within the Amazon rainforest's satellite imagery.

TAKEAWAYS

- In the preprocessing stage, the model analyzed label correlations through a co-occurrence matrix to understand inter-label relationships. A significant class imbalance identified in the multi-label analysis was addressed by applying calculated sample weights, ensuring equitable learning from all classes.

- The dataset was strategically split into training, validation, and testing subsets, creating a solid foundation for performance evaluation. Custom data generators optimized data handling for large datasets during model training and validation.

- The Convolutional Neural Network (CNN) model, tailored for various label classes, was fine-tuned to capture detailed patterns while preventing overfitting. The ensemble model, combining predictions from specialized CNNs, demonstrated high precision and a strong F1 score, indicating its capability to accurately classify complex features in satellite imagery and manage multi-label class imbalance efficiently.