Mapping Land Coverage Using Drone Imagery and Efficient Image Transformer Models

Aryaman Rtunjay *

March 25, 2024

Abstract

Solving environmental and anthropological problems requires effective public policy, which is only attainable when given accurate land use accounting. However, developing nations lack systems that can track these resources in real-time. Recent technological improvements have enabled Unmanned Aerial Vehicles (UAVs) to be deployed easily and inexpensively to image forests, farmland, and settlements. Manually processing such images to generate useful insights into environmental conditions and land use is time-consuming and expensive. Current automated tools utilize machine learning models like the Convolutional Neural Network, which require large dataset sizes. Nevertheless, collecting the required data is still slow and expensive. This study adapts and improves Vision Transformer models. Specifically, this study modifies the architecture of the Data Efficient Image Transformer (DEIT) to classify land cover images more accurately. This framework will provide developing nations with the accurate data and resources to create public policy. The optimized model reached a phenomenal F1-score of 95% and AUC of 99.85%. The outcome of this research is a precise machine learning-based framework that can produce accurate land coverage maps, a major stepping stone required for building a sustainable future.

1 Introduction

Developing nations are vexed with serious problems such as extreme population growth rates, land scarcity, unsustainable deforestation, extreme climate change, and unhealthy levels of air pollution [DHW+21, Ban23, But06, O'N23]. These problems are shadowed by those that impact citizens' quality of life. These nations' governments lack the appropriate data and resources to create effective public policy for solving these issues. Unless these problems are prioritized, most of humanity will live in these adverse conditions.

Sub-Saharan Africa is the world's fastest-growing region, predicted to reach a population of 2.2 billion people by 2050. This number is nearly 10 times

*Advised by: Dr. Parsa Akbari

relative to 1950 when only 227 million people populated this region [O'N23]. Sub-Saharan Africa also contains the most developing nations in the world. Out of the 46 Least Developed Countries (LDCs), this region contains 32 of them; this is nearly 70% of all LDCs [Nat22] (Figure 1). Sub-Saharan Africa also has a net loss of 1.5% of its trees each year, releasing approximately 460 kilotons of CO_2 [Agy98] This and other pollutants have devastated the air quality of Sub-Saharan Africa as shown in Table-1. Solving the issues in Sub-Saharan Africa will significantly reduce the number of countries with an LDC status and simultaneously improve the lives of millions, soon to be billions of people.

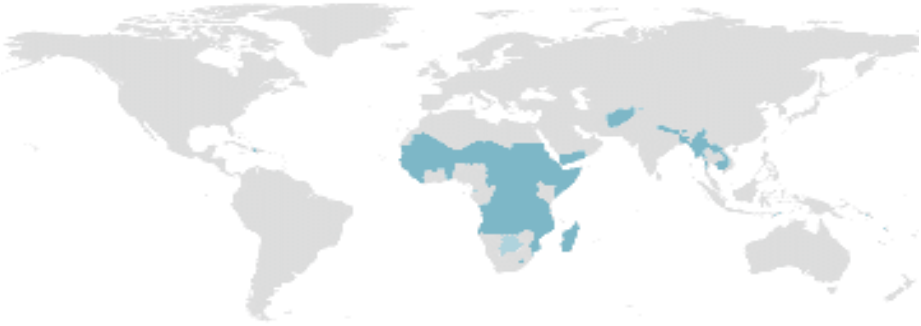

Figure 1: Shown in blue are the Least Developed Countries. The greatest amount of blue is in Sub-Saharan Africa [Nat22]

Rwanda was chosen as the country to study since it is a good representation of the condition of Sub-Saharan Africa, as described below. Rwanda has a population growth rate of 2.4%, which is over 2.5 times the global average of 0.9% [O'N23], but nearly the same as the population growth rate of Sub-Saharan Africa, at 2.59% [O'N23]. Rwanda has a net loss of 3% of its trees each year, leading to about 852 kilotons of carbon dioxide each year [DHW$^+$21]. Rwanda's deforestation rate is 3.75 times the global rate of 0.8% and is double that of Sub-Saharan Africa at 1.5% [Suz23, But06]. Table 1 shows the air quality of Kigali, the capital city of Rwanda, and it concludes that the air quality is extremely unhealthy. All of these factors lead to poor environmental and anthropological well-being. To solve these problems, it is necessary to have a solution that addresses both anthropological and environmental well-being simultaneously.

Most approaches have two problems: they only address anthropological or environmental well-being and require a tremendous amount of data, which is difficult to obtain in a developing nation. One approach to aid anthropological well-being performed an analysis of crop yield using machine learning. UAV imagery was used to track land cover, and a deep neural network was made to classify each image taken by the UAV. These results were then crossed with crop yield statistics over the past and used to provide information to farmers [TBLR23]. Another approach was done to slow urbanization. This approach involved land classification of urbanized areas and data from satellite telemetry. A CNN was used for the classification of these images to describe the type

Region	CO	CO_2	SO_2	PM 2.5
Rwanda	469	491	2.9	112
Sub-Saharan Africa	265	401	1.9	45
World	10	300	0.2	<15

Table 1: Shows the carbon e (CO) in parts per billion, carbon dioxide (CO_2) in parts per million, sulfur dioxide (SO_2) in parts per million, and Particulate Matter of size 2.5 microns (PM 2.5) in g/cm3 [Len23,Da22,Doc15,Bah23,Nps23, Ban23]. Rwanda and Sub-Saharan Africa are similar and much greater than UN Recommendations.

of structure or structures observed in the image and this was used to predict the rate of the city's growth and its impact on the environment [CRB+23]. As stated before, these projects focused on either anthropological well-being by measuring food scarcity or environmental well-being by predicting urbanization. Both projects also use neural networks that require tremendous amounts of data for accurate predictions which as previously mentioned is not possible in this scenario.

This study proposes an approach that can simultaneously help solve land scarcity, deforestation, and urbanization. This approach will employ mapping of crops, forests, and structures to show how Rwanda's cities, forests, and farmlands are changing over time. The data analysis will provide information on Rwanda's rate of urbanization, net deforestation, and farmland change. This visualization and information will help the Rwandan government make effective public policy, fast-tracking the course to solving issues in Rwanda and all other developing nations even past Sub-Saharan Africa. This study aims to improve anthropological and environmental well-being without requiring millions of data points.

This approach vastly improves on ideas introduced in prior works. The key improvements are the sheer preciseness of the mapping, the substantially reduced dataset size, and the variety of resources produced by this approach. This approach employs UAV data collection instead of satellites due to their considerably lower cost and highly detailed images [KCB+23]. This improvement allows for higher-quality data to be produced at only a fraction of the cost of prior approaches. One of the most significant improvements in the performance between mine and prior approaches is the 35% increase in F1-Score between the Legume class of a previous approach and this one. While the previous approach attained an F1-score of 45%, this approach soared above that figure, reaching a validation F1-score of 70% [CRB+23]. This improved predictive performance ensures maps are generated accurately, guaranteeing any public policy put in place will be effective. With the utilization of the improved DEIT model, this approach only requires 20% of the data previous approaches used. The benefit of such a reduction is reduced time and resources spent collecting data, thus

improving overall efficiency. Finally, this approach tracks farmland, forests, and structures, vastly more land cover types than prior approaches.

The main challenge encountered during the development of this approach lies in the model used to classify the images of the land cover types. Since UAV imagery was being used, a model was needed that capitalized on the detail of the images and was optimized to function with a small dataset. To overcome this barrier, experiments were conducted to determine the predictive performance of 3 models on this data. The first is the Convolutional Neural Network, the current State-of-the-Art model [DBK+23]. The second is Unlabelled Data Augmentation (UDA), which couples unsupervised machine learning with supervised machine learning. Third is the Vision Transformer, a new architecture primarily used in Natural Language Processing but becoming ubiquitous in Image Classification tasks [TCD+23]. Finally, the Data Efficient Image Transformer, an improved version of the traditional Vision Transformer meant to perform well when trained with fewer images [DBK+23]). The results of these experiments indicate that the Data Efficient Image Transformer outperforms its predecessors, the CNN and the Vision Transformer.

The study results show an average model validation F1-score of 95% and a model AUC of 99.85%, significantly surpassing the goal of 85%. The structure class has the lowest F1-score at 89%, possibly due to the low sample of data (166 images, average of 544 images across all classes). This would lead to a higher proportion of false positives, significantly affecting the F1 score. The AUC of this class (99.85%) further supports this theory; it shows that a small number of false positives makes a large false to true positive ratio, significantly affecting F1-score but not AUC.

2 Methodology

The proposed framework involves a multi-step approach to solving environmental and anthropological problems. The first step is to fly a UAV over crops, forests, and cities. The UAV will collect two data types: the image and its metadata. The metadata will be composed of the latitude/longitude coordinates of where the photo was taken and the picture's time stamp. The image data will be run through a machine learning model, which will classify the image into one of 4 categories, Banana, Maize, Forest, or Structure. This classification and the image data will be automatically mapped in real-time. Once classified, the image will be deleted to save storage space on the server, allowing the framework to be deployed for a significantly longer time. Figure 2 shows a visualization of this approach.

The procedure followed to accomplish the goals of this approach consists of data collection, preprocessing, model selection, model optimization, and model predictions. Finally, model classifications are used for map generation. Figure 3 shows a visualization of this process.

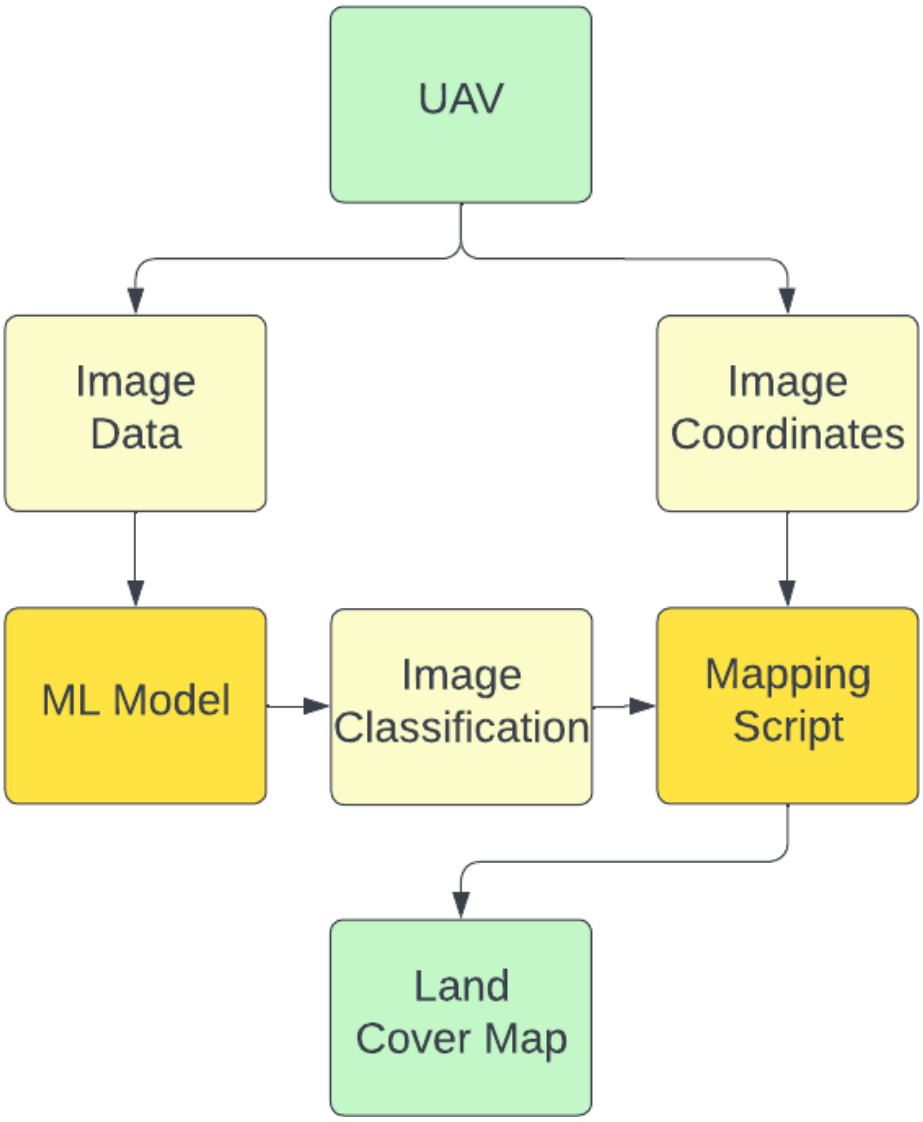

Figure 2: Flowchart visualizing overall approach framework. Light green refers to start/end, all yellows refer to intermediate steps. Darker yellow are key intermediate steps.

2.1 Data Collection

2.1.1 Data Description

The data are UAV images taken in 6 distinct agroecological zones within Rwanda. There were 19 land cover types, reduced to 3 crop types (Banana, Maize, and

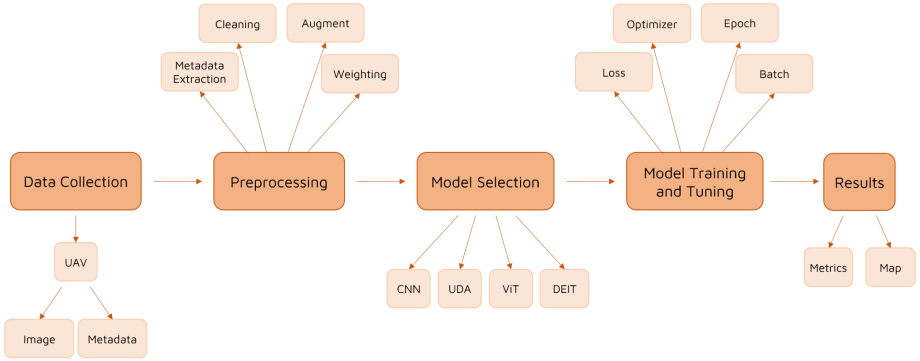

Figure 3: Procedure followed in constructing framework. Main events are shown in orange, sub-sections of the events are in light orange and have an arrow pointing to them.

Legumes), 2 non-crop types (Forest and Structure), and a catch-all Other type. Each data point contains an image, the latitude and longitude, and the land cover type. Each image has a P1, P2, or P3 designation corresponding to 3 different time periods. P1 is December 2018, P2 is January 2019, and P3 is February 2019 [CRB+23]. Figure 4 shows sample data from 4 classes. Figure 5 shows the position of each zone within Rwanda.

2.1.2 Data Source

Data was sourced from the publicly available RTI International Rwanda Crop Type Dataset (RTI Rwanda). RTI Rwanda used an eBee Plus UAV (Figure 6) to capture UAV images. The drone contained a GPS correction system allowing for a survey-grade accuracy of 10 cm without ground control points. This allowed for an efficient survey method using a relatively inexpensive drone to survey ground data. The images were taken with a resolution of 3 cm (each pixel represents 3 cm of land), for this, the drone flew at an altitude of 122 m above ground level [CRB+23].

2.2 Data Preprocessing

2.2.1 Data Cleaning

Many images had been corrupted, for example, some contained large black sections that removed a significant amount of information from the image. Other images were fully white across all classes making it nearly impossible to classify. Examples of such images are shown in Figure 7. This was due to an error when RTI Internation was transferring the image data into servers to be put on JSON files for their dataset [CRB+23].

Images were plotted on a 3-dimensional graph according to the average red,

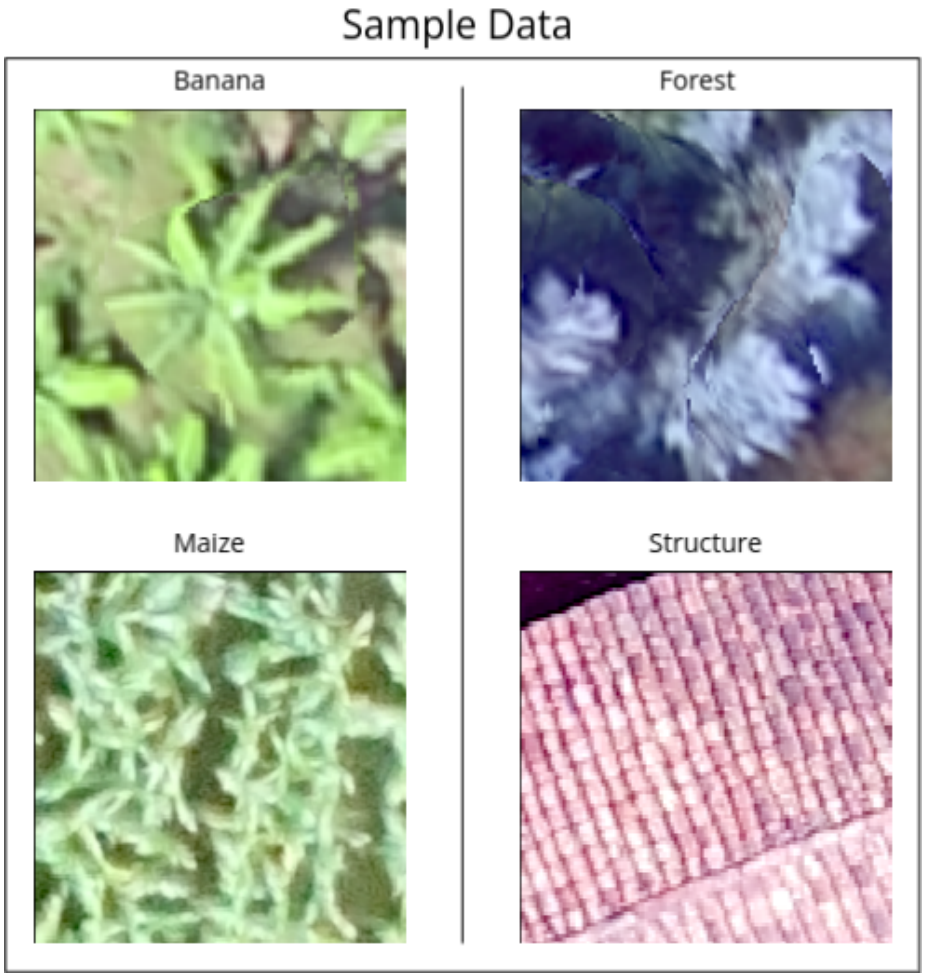

Figure 4: Examples of data for each class. Each image is high quality and shows great detail of each label, a trait of UAV Imagery.

green, and blue values across all pixels. Images that were not corrupted were found to follow a pattern, and a curve of best fit could be drawn for them. The graph was split into multiple subsections, and the average point for each subsection was found, by connecting each subsection a trend could be formed. By excluding points furthest away from the trend line, the corrupted images could be cleaned (Figure 8).

Figure 5: Map of Rwanda showing locations of each zone where data was collected.

(a) (b)

Figure 6: eBee Plus UAV, used to collect data in Rwanda. [CRB+23]

2.2.2 Data Augmentation

More variety must be introduced to the training data to prevent overfitting in the model. To accomplish this, 2 data augmentations were applied, Random horizontal flip and normalize. Random horizontal flip reflects the image using a vertical line at the center of the image. 50% of the images would have a horizontal flip applied. The values inside each image tensor were converted to be of a similar scale using the normalize() function. The images were converted from being in the range [0, 255] to the range [0, 1]. The mean and standard deviation used were the ImageNet default mean and ImageNet default standard deviation.

2.2.3 Data Weighting

Model performance was hindered by the lack of uniformity throughout the batch when creating batches of training and testing data. The banana and forest classes had the most data, so they had the greatest presence in the batch while classes such as maize had a minuscule presence. This led to low validation F1-Scores for the underrepresented classes. The solution to this problem is to ensure that all classes are equally represented in the batch. This was achieved by sampling the data and applying weights to each class. The total number of each image for each class was calculated, then a multiplier was created for each class to decide its presence in the batch based on the number of total data points it had. The formula to calculate the multiplier is

$$\frac{min(AllImages)}{ClassImages}$$

The denominator, Class_Images, is the number of images in the class for which the multiplier is calculated. The numerator, min(All_Images), is the minimum number of images in one class. This would mean that the class with the lowest images would have a multiplier of 1.0, while another class with more images would have a multiplier from $0.0 < \text{multiplier} < 1.0$ The WeightedRandomSampler() method was used to generate a list of probabilities for each data point in the dataset on the likeliness of it appearing in a batch. This list was the sampler for the DataLoader objects that would hold every batch.

2.3 Model Training

2.3.1 Model Selection

To create a machine learning architecture with the greatest predictive performance, it is crucial to compare multiple models. This study conducted a series of experiments with the goal of finding a machine-learning model that was optimal for use on the RTI crop-type dataset. The four models compared were the Convolutional Neural Network (CNN), Unlabelled Data Augmentation (UDA), the Vision Transformer (ViT), and the Data Efficient Image Transformer (DEIT). The CNN is regarded as the state-of-the-art model for image classification tasks,

however, its performance on a small dataset is largely limited [DBK+23]. UDA is an algorithm layered on top of the CNN, it involves training the CNN for both consistency along with accuracy, which, in theory, optimizes it for smaller data. This is done by creating an unlabelled dataset and an augmented unlabelled dataset. The CNN makes a prediction on both datasets and a loss function compares the difference between these predictions, training consistency. Accuracy is trained in the standard way, the CNN makes a prediction on a datapoint and a loss function compares the model prediction to the target value [XDH+23]. The Vision Transformer is a relatively new entry into image classification. This model employs an attention-based learning approach, where each image is broken into multiple segments. As the model trains, it identifies which segments are important to the image and which are not. As an example, if it sees brown in the section, it will understand that the section is soil and unimportant to the crop type in the image. This allows the model to quickly understand differences between images and train well on a small dataset [DBK+23]. The final tested model, the Data Efficient Image Transformer, is an improved version of the traditional Vision Transformer [TCD+23]. More details about DEIT are shown below in the 'Data Efficient Image Transformer' section. The results show that DEIT has the greatest performance of all tested models, further details about model testing and results are in the 'Experiments' section. Visualized below in Figure 10 is a bar graph showing the F1-scores of all 4 models. Table 2 shows the differences in the validation loss of each model.

	CNN	UDA	ViT	DEIT
Validation Loss	1.67	1.28	6.2e-03	2.2e-05

Table 2: Validation loss of each model, ViT performs significantly better than UDA and CNN. DEIT performs much better than all models.

2.3.2 Data Efficient Image Transformer

The transformer architecture has been ubiquitous in natural language processing tasks, and more recently, the Vision Transformer (ViT) architecture has been increasingly more prominent [DBK+23]. Vision Transformers break the image into different sections, the purpose of this being to find the areas of focus in the images to rapidly find key distinguishing features to classify. The benefit of using this model is that it requires significantly less data than the state-of-the-art CNN. DEIT is an expanded version of ViT that features a Distillation Token, allowing for an improved attention function over ViT, therefore needing fewer data and running faster [TCD+23] (Figure 9). Using models that require less data is necessary since UAV images are produced slower than satellite telemetry. The detail from UAV images also makes the transformer architecture's attention method even more effective than satellite telemetry.

2.3.3 Cost Function

The cost function measures the difference between the model's predictions and the target value. This shows the model's inability to find a relation between its inputs and outputs [Koe23]. The cost function of a neural network is defined as the sum of the errors in each network layer. The cost function used in this project is Cross Entropy Loss. The loss of a model is the difference between the softmax value of the predictions for each class and the target output for that input [Koe23]. The variable's entropy is the uncertainty inherent in the possible outcomes or the probability of making a correct guess at random [Koe23]. The entropy and the loss are then crossed with each other to compute a final loss which is how inaccurate the model is; a perfect model will get a 0. Cross Entropy Loss was chosen as the cost function since it is designed to be used for models that are performing multi-class predictions. A modification was made to Cross Entropy Loss in a customized loss function made in this approach. The model typically confused Maize and Banana, reducing the metrics for these classes. To account for this, a scheduler was created. This scheduler was made using a value added to the loss to increase it if there was a maize/banana misclassification. The added value changes based on how many steps the model has trained for; the value decays as the model trains more. The final loss ($Loss_f$) is calculated by adding the curved value, which is inversely proportional to the number of steps (x), to the initial loss ($Loss_i$). The formula and curve used are shown below:

$$Loss_f = Loss_i + \frac{4^3}{x^{0.65}} + 0.01$$

2.3.4 Optimizer

The purpose of the optimizer is to adjust the learning rate and the model weights to maximize model performance. The optimizer chosen was the ADAM (Adaptive Movement Estimation) Optimizer. The ADAM optimizer is a first-order gradient-based optimization algorithm for a function [KB23]. A first-order method is defined as one that only requires the first derivative/gradient of the model performance function. The ADAM optimizer was chosen because it has a faster computation time and fewer parameters that require tuning [Par23]. The parameters tuned were the learning rate, which was set to $4 \cdot 10^{-7}$, and the weight decay was set to $1 \cdot 10^{-7}$.

2.4 Model Evaluation

All metrics are calculated using the model's true positives (TP), False Positives (FP), True Negatives (TN), or False Negative (FN). The 4 metrics used to evaluate model validation performance were precision, recall, F1-score, and AUC. Precision aims to answer the question of what proportion of positive identifications was correct, or the percent of true positives identified [Boa23]. The

formula to calculate precision is

$$Precision = \frac{TP}{TP + FP}$$

The recall is meant to identify what proportion of actual positives were identified correctly [Boa23]. The formula to calculate recall is

$$Recall = \frac{TP}{TP + FN}$$

Precision and recall are typically in tension with one another; improving one will reduce the accuracy of the other, so to find the true performance of the model, the F1-Score is employed. The F1-score is the harmonic mean of the model's precision and recall [Boa23]. The formula to calculate the F1-Score is

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

A perfect model will have an F1-score of 1, meaning a precision and recall of 1. The F1 score is the most applicable metric for this project since it prioritizes positive predictions. Having a false negative in this scenario does not make as much of an impact as having a false positive. The main downside of the F1-score is its inaccuracy in depicting model performance in classes with a small number of samples compared to other classes. The F1-score would focus on a higher false-to-true positive ratio and not on the true and false positive rates, which is why AUC is used. AUC is the area under the ROC curve, a graph where the x-axis is the False Positive Rate and the y-axis is the True Positive Rate. The formula for the True Positive Rate is equivalent to the formula to calculate recall. The formula to calculate False Positive Rate is

$$FPR = \frac{FP}{TN + FP}$$

and the formula to calculate AUC is

$$AUC = \int_0^1 r(F)\, dF$$

where r is the ROC curve function and F is the false positive rate.

3 Experiments

3.1 Equal Weighting of Supervised and Unsupervised components increases UDA F1-Score by 33%

Unlabelled Data Augmentation (UDA) allows unsupervised data to improve supervised model performance. This is done by training for consistency with an unlabelled and unlabelled augmented dataset, and training for accuracy with

the supervised dataset. The UDA cost function includes a supervised and unsupervised component weighted by lambda [XDH+23]. The hyperparameter was incremented from 0.0 to 0.9 to identify the optimal lambda. A lambda of 0.5 led to a Validation F1-Score improvement of 33% and the validation loss by 0.20 on average. Thus, optimal model performance is reached when the supervised and unsupervised components are weighted equally. Figure 11 shows a plot of the results of this experiment.

3.2 Vision Transformer outperforms CNN with F1-Score increase of 460%

The transformer architecture involves breaking the input into multiple sections and then identifying the most important sections, resulting in increased attention to those sections. The transformer architecture is most commonly used on Natural Language Processing problems, but more recently, its presence has been widespread in image classification tasks [DBK+23]. It's been shown in prior studies to outperform the previous state-art-of-the-art model, the CNN, in low-data situations [DBK+23, TCD+23]. The training and validation loops were kept the same, the only difference being the model used. The Vision Transformer tremendously outperformed the CNN, with a 460% increase in validation F1-score and a validation loss improvement of 1.67. This confirms prior research, which shows that the Vision Transformer Architecture outperforms the CNN in low-data scenarios. The final results of this experiment are shown in Table 3.

Model Used	F1-B	F1-F	F1-M	F1-S	F1-Avg
ViT	87%	91%	89%	55%	73%
CNN	5%	19%	30%	11%	13%

Table 3: Results of Experiment 2. F1-B is F1-Banana, F1-F is F1-Forest, F1-M is F1-Maize, F1-S is F1-Structure, and F1-Avg is F1-Average. Results show ViT significantly outperforms the traditional CNN.

3.3 Data Efficient Image Transformers outperform Vision Transformers with F1-Score increase of 3%

The Data Efficient Image Transformer is an improved version of the traditional Vision Transformer. The key difference between the two is a distillation token from a pre-trained model, allowing transfer learning to aid model performance. DEIT is shown to improve over the traditional architecture in low-data scenarios slightly [TCD+23]. This experiment shows that the DEIT architecture does show an improvement in model performance, with a validation F1-score increase of 3% and a validation loss that was an order of magnitude superior. As outlined in previous studies, this study has also found that the DEIT architecture

outperforms the ViT in low-data scenarios. Table 4 summarizes the results of this experiment.

Model Used	F1-B	F1-F	F1-M	F1-S	F1-Avg
DEIT	87%	91%	89%	55%	73%
ViT	83%	88%	86%	58%	71%

Table 4: Columns same as Experiment 2. Results of Experiment 3 show that DEIT outperforms ViT, thus also outperforming the CNN.

3.4 Images containing intercropping reduce model performance with F1-Score increase of 9%

Farmers tend to grow some types of crops with others, a practice known as intercropping [CRB+23]. In this case, images of legumes consistently contained either a banana crop or a maize crop in the same image [CRB+23]. A test was conducted to see whether the inclusion of the legume category was causing model performance to deteriorate for both the banana and maize classes. The experiment results show an average validation F1-score increase of 9% and a validation loss improved by 2 orders of magnitude when removing intercropped images (Table 5). This points to the conclusion that intercropped images lead to the model inaccurately predicting which crop is the correct one to predict when seeing intercropped images, and the removal of which removes this confusion.

Intercropping	F1-B	F1-F	F1-M	F1-S	F1-Avg
Present	87%	91%	89%	55%	73%
Not Present	97%	96%	96%	89%	95%

Table 5: Columns same as Experiment 2 and 3. Results of the experiment show model is confused from the presence of intercropped data, causing lowered predictive performance.

4 Results

4.1 Model Results

This framework aimed to design an efficient and inexpensive method for developing nations to map land coverage to create effective public policy. The performance goals were to achieve a model validation F1-score and AUC of 85%. This approach achieves a model validation F1-score of 95% and an AUC of 99.85%. In addition, the model precision was 97% and the recall was 93%.

This is achieved given a small amount of training data, reducing the cost of applying this framework in a developing nation since less time and resources must be devoted to flying a UAV over land and photographing it. The Receiver Operating Characteristic (ROC) curve for each label shows the model is an exceptional classifier, with the AUC being 1.00 for 3 classes and 0.99 for the fourth. The confusion matrix also shows this since most predictions lie on the diagonal, meaning the model predicted almost all images correctly. The ROC Curves are shown in Figure 12 and the confusion matrix is shown in Figure 13.

4.2 Land Cover Maps

After ensuring the predictive performance of the model was exceeding requirements, it was utilized to create land cover maps of each label. Each datapoint was coupled with its coordinate and when the model classified a datapoint, it could easily be associated with the proper coordinate. The maps created showed a clear distinction between the types of crops that were being grown, where each forest was and its size, and the position of each structure throughout the 6 zones. The high AUC and F1-Score of the model ensure that the maps created are highly accurate. Refer to Figure 14 for the final land cover maps.

5 Discussion

This study has investigated the performance of different deep learning architectures and techniques in the context of crop classification using remote sensing data and has revealed some important findings. Firstly, the use of UDA can significantly improve the performance of supervised models, with a lambda of 0.5 leading to the greatest model performance. Secondly, the ViT architecture has outperformed the CNN architecture in low-data, high-detail scenarios, and the DEIT Architecture performs even better. The third finding is the inclusion of intercropped images in the dataset deteriorating model performance, and removing them can significantly improve results.

These findings have significant implications for the field of crop classification using remote sensing data. The use of UDA can lead to more accurate crop mapping and monitoring. The ViT and DEIT architectures can provide better accuracy than traditional CNN models, resulting in more precise crop mapping as well. The removal of intercropped images leading to an increase in model performance suggests that intercropped images create noise in the dataset.

The results of this study are consistent with other studies that use the transformer architecture in the field of remote sensing and crop classification. Two studies exist that perform similar classifications, one identifying weeds vs crops and another detecting plant disease. Both studies employ ViT and both show that ViT significantly outperforms CNNs in their classification [RDCH23, TBLR23].

This study has a few limitations that future work can address. The dataset used in the study is limited to two crop types, and future studies can explore

other crops and land-use types. Additionally, the study only considers the impact of intercropping on model performance, while other factors such as cloud cover and atmospheric effects can also affect the accuracy of remote sensing data [KCB+23]. Finally, the study could benefit from adjustments to either the model or the data to allow the model performance to not be affected by the inclusion of intercropped data. Intercropping is a common practice in some farms so its exclusion is not the best method to solve this issue [CRB+23]. Future studies could employ a custom loss function, or reclassify intercropped images to include the names of all crops present in the image.

Overall, this approach is a novel and effective solution to better the conditions of developing nations. The creation of real-time land cover maps will aid in the creation of public policy at a minimal cost. This means the study has the potential to stop 880 kilotons of CO_2 emitted each year in Rwanda through deforestation, slow the rate of urbanization in Rwanda down to the global average, and allows for the optimal placement of farmland, ensuring not only no more land scarcity but also food security.

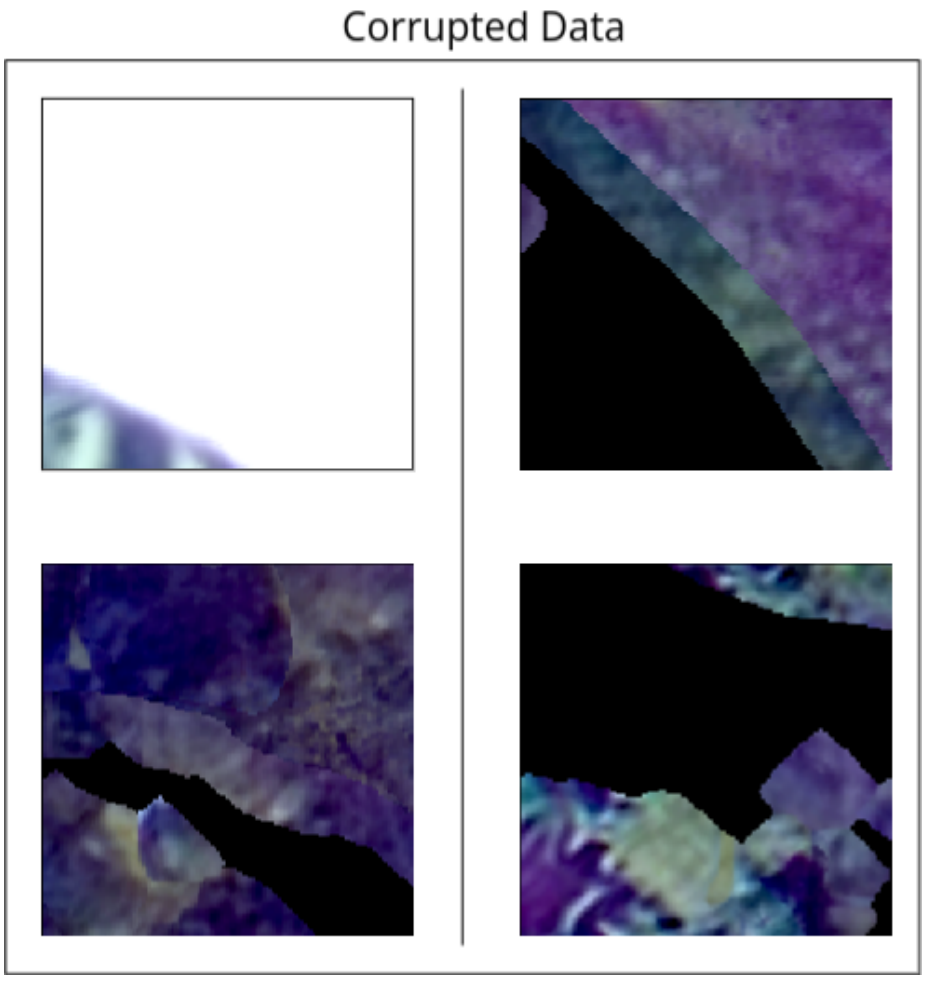

Figure 7: Examples of corrupted data. Problems are always excess black (RGB: 0, 0, 0) and excess white (RGB: 255, 255, 255).

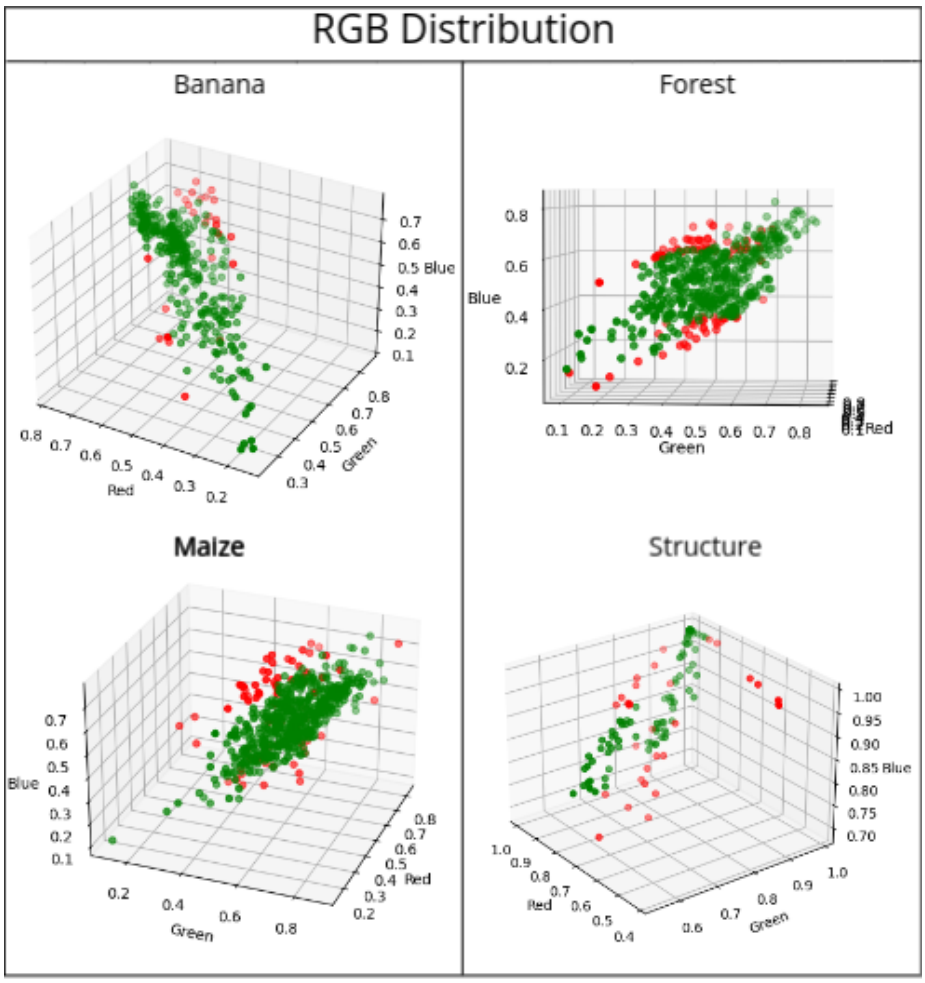

Figure 8: Average R, G, and B graphed for each label, the curve of best fit drawn, and points sufficiently far from the curve marked as red and removed.

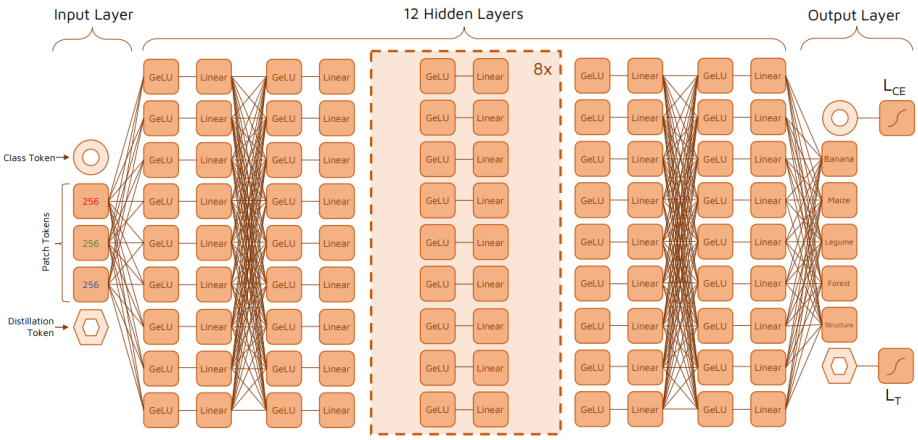

Figure 9: Improved DEIT Architecture utilized in this study. There are 12 hidden layers, each with 9 Gaussian Linear Unit 'identity' neurons and 9 linear 'feed-forward' neurons. Cross Entropy Loss and Teacher loss are computed after the output layer. [TCD+23].

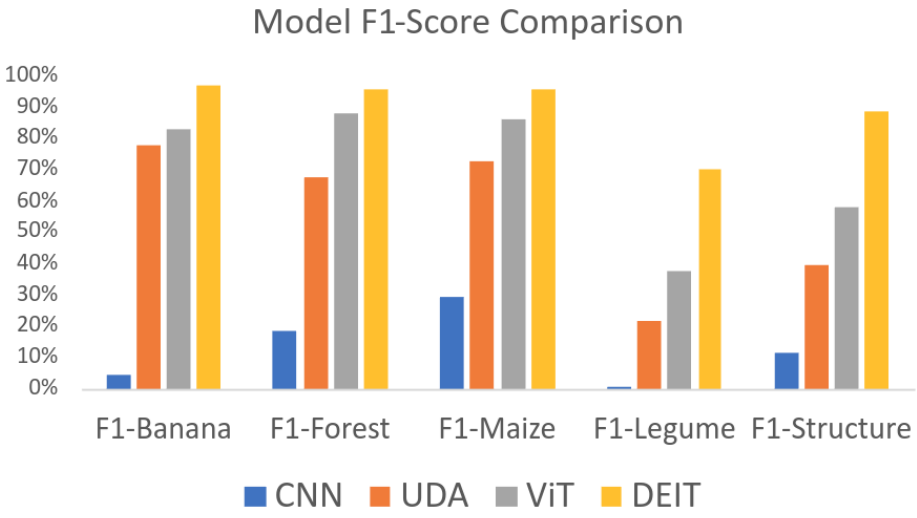

Figure 10: F1-scores of the 4 tested models, DEIT (yellow) consistently performs best. Also shown is that CNN performs extremely poorly with less data.

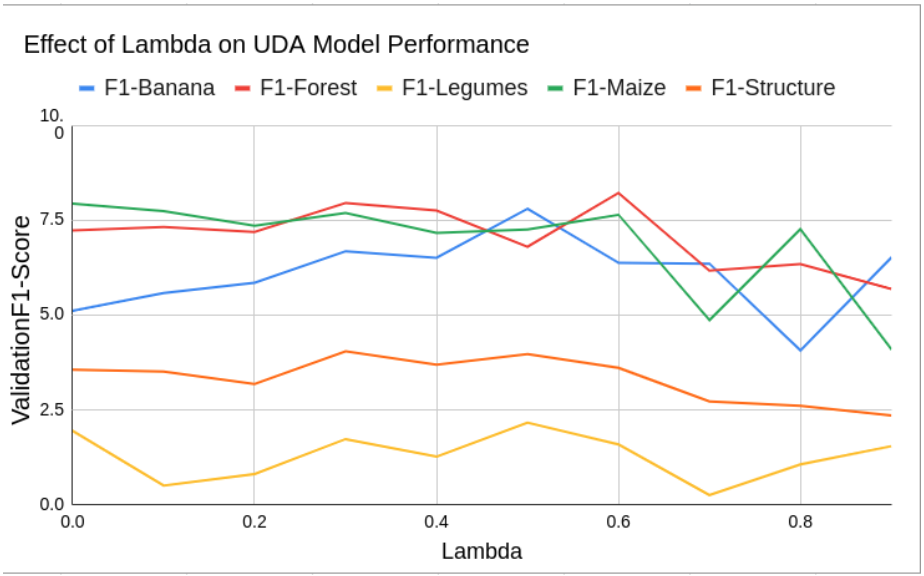

Figure 11: UDA per label F1-score trend as Lambda changes. The F1-Score across all labels is at max or very high at a lambda of 0.5

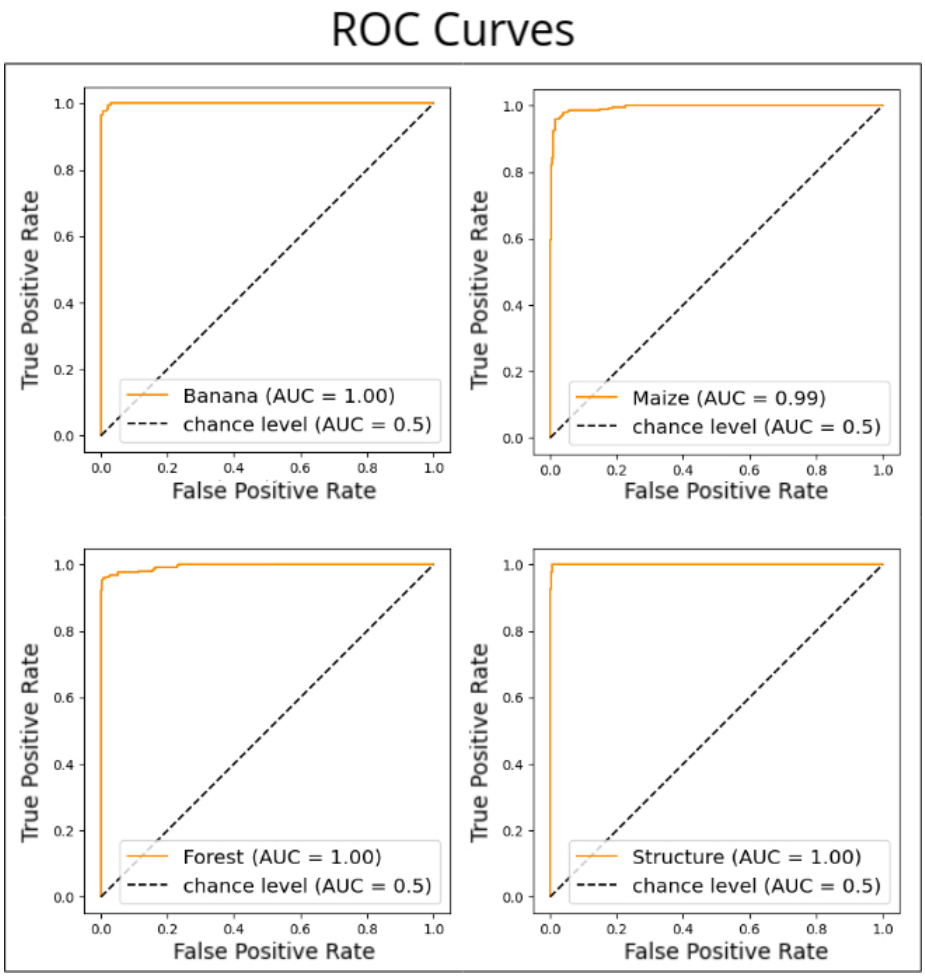

Figure 12: Receiver Operator Characteristic (ROC) Curves for each class.

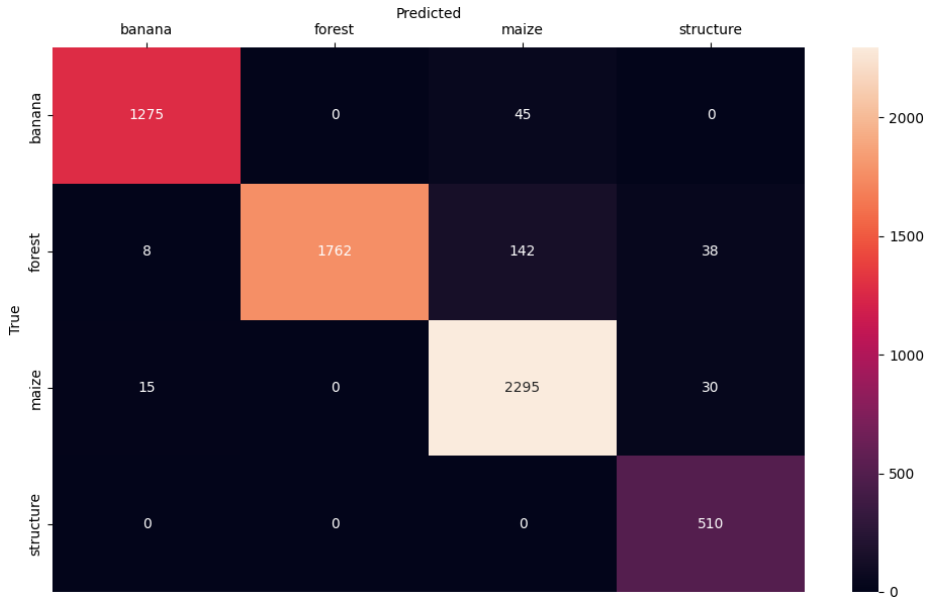

Figure 13: Confusion Matrix for model predictions, predicted labels on x-axis, true labels on y-axis.

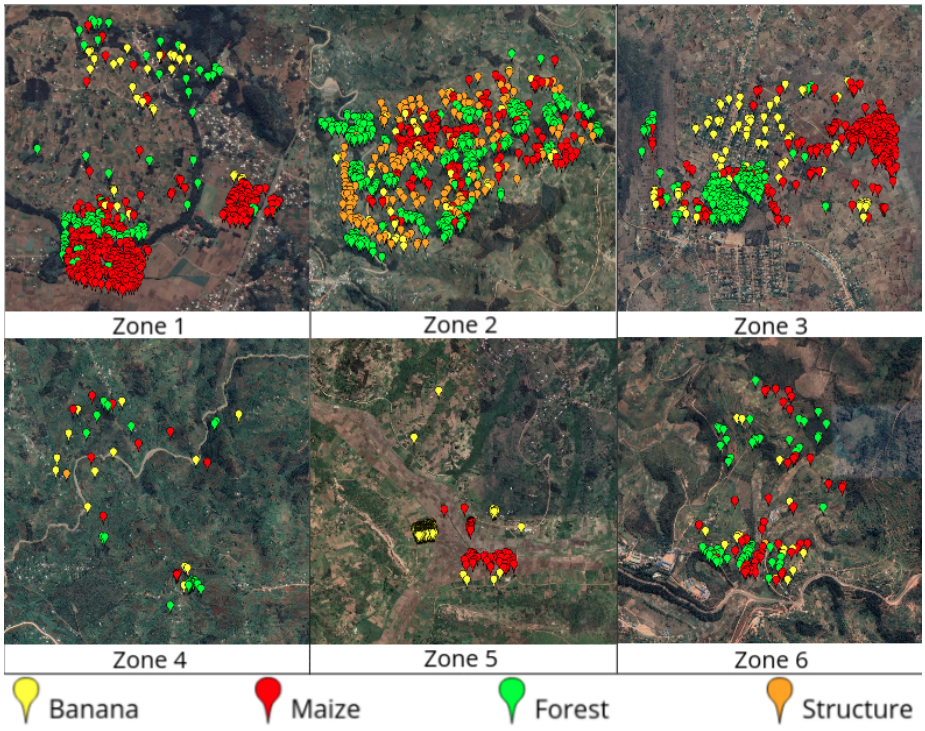

Figure 14: Maps of model predictions of each agro-ecological zone.

6 Acknowledgements

I would like to profusely thank my mentor, Dr. Parsa Akbari, research associate at the University of Cambridge, for offering his advice, guiding me through my project, and having meaningful discussions with me throughout!

References

[Agy98] Y. Agyei. Deforestation in sub-saharan africa. *Deforestation in Sub-Saharan africa. Retrieved March*, 1998.

[An22] U. S. An. Aqi basics. *AQI Basics — AirNow.gov. Retrieved March*, 2022.

[Bah23] V. (2018 Baharane. August). *EVALUATION OF THE VARIABILITY OF SULFUR DIOXIDE, CARBON MONOXIDE AND CARBON DIOXIDE CONCENTRATIONS IN KIGALI CITY*, March 2023.

[Ban23] R. Bansal. Kigali air quality index (aqi) : Real-time air pollution: Rwanda. March 2023.

[Boa23] S. (2022) Boaz. Multi-class metrics made simple, part ii: The f1-score. March 2023.

[But06] R. Butler. Sub-saharan africa deforestation. 2006.

[CRB+23] R. Chew, J. Rineer, R. Beach, M. O Neil, N. Ujeneza, D. Lapidus, T. Miano, M. Hegarty-Craver, J. Polly, and D. S. (2020) Temple. Deep neural networks and transfer learning for food crop identification in uav images. March 2023.

[Da22] U. S. Da. Carbon dioxide health hazard information sheet - food safety and inspection service. *United States Department of Agriculture*, March 2022.

[DBK+23] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. (2021 Houlsby. June 3). *An image is worth 16x16 Pixels: Transformers for image recognition at scale*, March 2023.

[DHW+21] C. Davis, N. Harris, M. Weisse, E. Goldman, D. Gibbs, J. Webb, and L. Amaral. Rwanda deforestation rates & statistics: Gfw. *Global Forest Watch*, March 2021.

[Doc15] E. Docket. Proposed rule to implement the fine particle national ambient air quality standards. *EPA*, March 2015.

[Gov23] R. (2019 Government. Rwanda environment and climate change analysis. pages 2019–06, March 2023.

[KB23] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. March 2023.

[KCB+23] A. Khaliq, L. Comba, A. Biglia, Ricauda Aimonino, and Chiaberge D. Comparison of satellite and uav-based multispectral imagery for vineyard variability assessment. March 2023.

[Koe23] K. E. (2022 Koech. Cross-entropy loss function. March 2023.

[Len23] T. Lenander. Co-limit. March 2023.

[Nat22] U. Nations. World population prospects. *United Nations*, 10:2023, March 2022.

[Nps23] U. S. Nps. Sulfur dioxide effects on health. *National Parks Service*, March 2023.

[O'N23] A. O'Neill. January 27). *Sub-Saharan Africa - population growth*, March 2023.

[Par23] S. (2021 Park. A 2021 guide to improving cnns optimizers adam vs sgd. March 2023.

[PBMS23] T. R. H. Pearson, S. Brown, L. Murray, and G. Sidman. Greenhouse gas emissions from tropical forest degradation: An underestimated source - carbon balance and management. March 2023.

[RDCH23] R. Reedha, E. Dericquebourg, R. Canals, and A. Hafiane. Vision transformers for weeds and crops classification of high resolution uav images. arxiv.org. 2023.

[Rot23] K. (2022 Roth. January 13). *World Report*, 2022, March 2023.

[Suz23] E. Suzuki. World's population will continue to grow and will reach nearly 10 billion by 2050. March 2023.

[TBLR23] D. Temple, R. Beach, D. Lapidus, and J. Rineer. Using satellite images and artificial intelligence to improve agricultural resilience. 2023.

[TCD+23] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou. Training data-efficient image transformers & distillation through attention. March 2023.

[XDH+23] Q. Xie, Z. Dai, E. Hovy, M.-T. Luong, and Q. Le. Unsupervised data augmentation for consistency training. 2023.