Uniting Machine Learning and Citizen Science for Automatic Land Cover Classification

Oseremen Ojiefoh,[*†] Roayba Adhi,[‡] Riya Tyagi,[¶] Haitham Ahmad,[§] Lavanya Gnanakumar,[‖] and Kavya Ram[⊥]

†Rick Reedy High School, Frisco, TX 75036, USA

‡Washington-Liberty High School, Arlington, VA 22203, USA

¶Phillips Exeter Academy, Exeter, NH 07078, USA

§Osbourn Park High School, Manassas, VA 20111, USA

‖Folsom High School, Folsom, CA 95630, USA

⊥Allen High School, Allen, TX 75002, USA

*Corresponding author. Email: seeszooniverse@gmail.com

Abstract

To complement satellite land cover data, researchers today rely on citizen science to record and analyze ground photos of land cover because of its ability to efficiently facilitate research at a large scale, quickly gather data across a variety of regions, and ethically engage communities in pertinent matters. The GLOBE Observer App, part of The GLOBE Program, is a citizen science tool used to document the planet. To make a land cover observation, a user logs their date and location, records the surface conditions of the observation area, and takes six directional photos. The land cover features present are then identified in each photo, with users dragging sliders to report the percentage of land or sky that each feature covers in the image. However, our surveys have found that the classification step can be a complicated process. User subjectivity and limited knowledge of land cover terminology, paired with the time-consuming nature of classification, caused 46% of survey-takers to report that they often or always skip the step of classifying altogether, substantially decreasing data usability for scientists. In this study, we contributed to developing a geographically diverse dataset of 5,896 directional land cover photos, which were then uploaded to the GLOBE database. Using Zooniverse, a citizen science annotation tool, labels determining whether the three classes—sky, land, or water—were present in each land cover photo were collected. Three ResNet-18 machine learning models were trained to classify sky, land, and water in each photo, treating Zooniverse labels as ground truth. The sky, land, and water models received an accuracy of 94%, 92%, and 95%, respectively. The trained models were then used for feature extraction. Features were passed into three Support Vector Machines, which determined the percentages of sky, land, and water in each photo. Our model pipeline can assist untrained volunteers with future image classifications, resulting in a higher-quality dataset for scientists.

Keywords: land cover; citizen science; machine learning

1. Introduction

Over the last decade, citizen science has gained significant prominence as a powerful tool for streamlining research, advancing scientific literacy, and promoting learning through active engagement practices (Frigerio et al., 2021). In the field of land cover, citizen science can be a valuable way to collect data on a large scale, supplementing studies of Earth's vast terrain and how it is being used (Kohl et al., 2021).

1.1 Introduction of the GLOBE Observer

One citizen science tool transforming land cover is the Global Learning and Observations to Benefit the Environment (GLOBE) Observer app, part of The GLOBE Program. This application utilizes citizen scientists to photograph and classify land cover through its Land Cover: Adopt a Pixel tool. For optimal documentation, users are advised to conduct observations in areas that are approximately the size of a soccer field. Each observation consists of five main steps: logging the date and location, recording the presence or absence of certain surface conditions (snow/ice, standing water, muddy ground, dry ground, leaves on trees, and rain/snow), capturing six directional photos (Up, Down, North, East, South, West), selecting the land cover features present in each North, East, South, and West (NESW) photo, and identifying what percent of the ground or sky each land cover feature covers. Volunteers' observations can then be used to conduct research at spatial resolutions that cannot be achieved through the use of satellites alone (The GLOBE Program, n.d.).

Figure 1. (a) Screenshot of feature identification step; (b) screenshot of percentage selection step.

1.2 Limitations of GLOBE Observer

Surveys conducted by our team have found that the step of classifying observation photos can be complex and time-consuming for untrained volunteers. When surveyed, 73% of respondents claimed that they were not confident in their labeling[1], with 82% finding it difficult to navigate the labeling process[2]. Moreover, 46% had been dissuaded from labeling the images altogether[3] due to the length of the process and complications that arose during the classification stage. Prior to collecting data, users are prompted to complete a comprehensive in-app training that covers the definitions of land

[1]% of respondents who answered <5 on a scale from 1-10 when asked, "How confident are you in your labeling?".

[2]% of respondents who answered <5 on a scale from 1-10 when asked, "How difficult was it to navigate the app's labeling process?".

[3]% of respondents who answered <5 on a scale from 1-10 when asked, "How often did the length/complications of the process dissuade you from labeling the images at all?".

cover features, how to capture the directional photos, and the general protocol to follow when labeling the images. Furthermore, users can only receive the 'Land Cover Expert' Badge if they complete additional training that prompts the user to classify sample images until they match the app's 'correct' classification (including the land cover features and their individual percentages of land or sky cover) (Kohl et al., 2021). However, although users are provided this training and must review their observations before uploading to the GLOBE database, there are few safeguards in place to prevent the submission of incorrect classification data. The lack of a consistent labeling methodology as well as a filtration system for faulty images (beyond those that go against GLOBE guidelines) or classifications, can produce an unreliable and incomparable data set to be used for future studies and analyses (Manzanarez et al., 2022).

Figure 2. Results from survey regarding GLOBE Observer app user satisfaction.

1.3 Our Proposed Solution

To provide a solution that can decrease erroneous classifications and increase the quality of land cover classifications submitted, our research explores both the feasibility of using volunteer-labeled data in training a 'land cover feature suggestion' model, and the efficacy of said model in identifying three feature categories relevant to the GLOBE Observer app.

1.4 Why Use Machine Learning?

ML models are a prominent analytical tool employed to track, delineate, and assess changes in land cover over time. Using remote sensing data, these models can be trained to accurately classify land cover found at various spatial scales, ultimately improving classification performance with less uncertainty (Yuh et al., 2023). Furthermore, ML models are preferred due to their simplicity, flexibility, speed, and cost-effectiveness in comparison to other artificial intelligence methods (Alshari et al., 2023).

1.5 Previous Research Done on Topic

Several deep-learning models have already been trained to identify land cover. Related research has been done using GLOBE data sets, with Manzanarez et al., 2022 focusing on aggregating feature labels across images (North, South, East, West, and Up), rather than the serving composition of one. Additionally, the Places365 data set, developed by Zhou et al., 2018, is a data set curated to train convolutional neural networks (CNNs) on scene recognition. Although this project provides a pre-trained ResNet model for image classification, the categories did not align with the features recognized in the GLOBE Observer app. Moreover, the model outputs the probability of each image being a specific scene, not the percentage that a land cover feature composed in an image.

1.6 Purpose of Our Study

The goal of this study was to use data acquired and labeled by citizens to train three separate deep-learning models (ResNet-18). These models were employed to analyze land cover images, and output whether or not an image contained sky, land, or water. Due to time and resource constraints, this study is the first step in developing a model that can identify all land cover features available in the GLOBE Observer app, along with their relevant percentages (% of ground covered or % of sky covered). Our research attempted to answer two main questions: are citizen scientists an efficient way to process data to train an image segmentation model, and how effective is this model in making our desired predictions?

1.7 Significance

By providing users with suggestions on what might be in their photo, our proposed solution would enhance the quantity and quality of submitted data by reducing the time and uncertainty that comes with making classifications.

2. Methods

2.1 NASA SEES 2023 Data set

When preparing our data set, various sources were considered for training images. It was ultimately decided to train our model off of fellow SEES 2023 Interns' Land Cover Observation photos. These photos were chosen for two reasons: explicit permission was given to use the data and the interns' photos matched the format, consistency, and quality of the images our model was targeted to classify.

The team, titled SEES2023 on the GLOBE Observer app, consisted of 54 members and made a total of 1,477 Land Cover observations. A majority of the team's observations were made in each member's Area of Interest (AOI), a 3km x 3km area of land. Each AOI was then further divided into 37 squares that were each 100m x 100m squares of land. A land cover observation was to be made at each of these points, which consisted of six images: North, East, South, West, Up, and Down. Across the 46 AOIs used, there were 8,862 total images and 5,908 NESW photos. These observations were taken from over a dozen states in the United States and 3 continents (North America, South America, and Asia). After filtering out unusable photos, our data set consisted of 5,896 photos.

Scripts written in Python were then used to organize the data from each photo found in the GLOBE API. Each image's link was then extracted and used to download the images onto a shared Google Drive folder. These images were uploaded to Zooniverse, another citizen science tool, for classification. On the GLOBE Observer App, there are eight main land cover classes, which then branch off into 49 specific categories across each class. In the app, users indicate the percent area of each land cover category in an image. However, in our Zooniverse workflow, volunteers only indicated whether an image contained one or more of these categories: sky, land, and water. Using the aid of 11 citizen scientists, our 5,896 subjects were completed and we received 5,996 classifications due to the double classification of some images. Thus, the repeated classifications were filtered out, resulting in a data set of 5,896 NSEW photos with volunteer-assigned labels. These labels were then refined into binary classifications for each image that described the presence of each of the three categories.

2.2 Labeling Images on GLOBE Observer

water	sky	land	img_name
0	0	1	3482763.jpg
0	1	1	3482847.jpg
1	1	1	3422256.jpg

Figure 3. Sample Images and their classifications.

2.2.1 Tutorial

To begin the process of recording land cover observations in the GLOBE Observer App, users click on the application and are met with a screen that displays the four tools, or protocols, that are available to record observation: Clouds, Mosquito Habitat Mapper, Land Cover, and Trees. From there, users click on the Land Cover: Adopt a Pixel tool, and are given a tutorial to read through. The tutorial begins with an introduction to the GLOBE Observer Land Cover App and its purpose, as well as the significance of the data that is to be collected with this tool. Following the Introduction, there is a section titled, "What Do I Do?", which consists of an animation that displays the main steps of taking the land cover observations. After the user completes the required steps of logging an adequate observation area, and capturing photos of the land cover in the area, users are shown three optional steps: selecting the land cover features seen in the NESW photos, estimating the percentage of each land cover feature covers of the ground (or the sky for trees), and comparing their Modified UNESCO Classifications of their area to those recorded by satellites. After this section, there are pages that describe when and where observations should be taken in greater detail. The tutorial then ends after a final page that offers more information about the GLOBE Observer app and The GLOBE Program.

2.2.2 Capturing Images and Recording Field Notes

Once the tutorial is completed, the user is taken to a page that allows them to see the number of observations they have submitted to the GLOBE database and the badges they have earned from completing the in-app trainings and recording observations. Underneath this information, the user is able to make a new observation or review/send observations saved in the app.

If the user chooses to make a new observation, they are shown a caution page that describes how to stay safe when taking photos, but this can be turned off if the user so indicates. The user then enters the date and time of the observation and then verifies the location. There is the option of moving around a red marker to correct the latitude and longitude detected by the app originally, and users can reset the estimated location accuracy if needed. The user then indicates the presence of the following surface conditions: Snow/Ice, Standing Water, Muddy, Dry Ground, Leaves on Trees, and Raining/Snowing by either selecting Yes or No. Next, users are advised to turn their device to landscape mode to take Up and Down photos, and then NESW photos, in which onscreen directions are available. The options of taking the pictures manually or uploading images from one's camera roll are also available. Once the photos are taken, they are shown above the Next button. The photos can be retaken, replaced with one from the user's camera roll, or removed altogether if the user wishes as well.

After the images have been taken and finalized, the user has the option of recording field notes to provide additional information about the location. These notes may only be edited before the observations are submitted to the GLOBE database.

Upon completion of the Field Notes section, the user is once again presented with all six of their directional photos and is reminded of what they should and should not include. At this stage, before choosing to analyze the land cover or skipping the analysis and saving the photos, the user may replace any of the six directional photos if they wish. If the user chooses to skip the analysis stage, they are presented with all of the data they have inputted for this observation which includes the time and date, location, NESW and Up and Down photos, and their field notes.

However, if the user chooses to continue to analyze the land cover in the images further, a few more steps must be taken before submission. As mentioned prior, the main focus of this project was oriented around this stage of the observations, due to a lack of a consistent methodology for classification, as well as its tedious and time-consuming nature. Moreover, with the absence of a quality-assurance stage, there are few barriers to prevent the submission of inaccurate classifications, thus there was a need to address this issue.

2.2.3 Image Classification

To begin the image classification, users first select the option to continue to analyze the land cover in their NESW photos on the Completed Photographs stage. Upon selecting this option, users are presented with a preview of their North photo in the top left-hand corner, which can be enlarged by clicking on the image. On the right-hand corner of the screen, next to the writing that describes which directional photo is shown, there is a Next button and a counter that displays how many land cover features have been selected.

Underneath each of these app features, there are drop-down menus for the following land cover classes: Trees, Shrubs, Herbaceous/Grassland, Barren, Wetlands, Open Water, Cultivated, and Urban. If the user chooses one of these classes, the drop-down menus display more specific options that fall under the general land cover category, with each of the features assigned a distinct MUC code (see Figure 3 for all of the land cover features recognized in the GLOBE Observer app). On the right-hand side of each land cover option, there is a button that may be pressed to provide definitions and images of each land cover feature. Additionally, every time the user selects a land cover feature, a small square corresponding to the color of the feature appears on the right-hand side of the larger drop-down menu, next to the information button.

Once the user has selected all of their options, they are taken to the Select Percentage screen. On the top of the screen, the user is able to select the previews of the Up, Down, and directional photo (whichever one of NESW that may be) to enlarge them. Underneath these previews, there are sliders for each of the land cover features selected to indicate the percentage of ground or sky (only for the Trees class) they cover in the area. The user may delete any of the land cover features or add

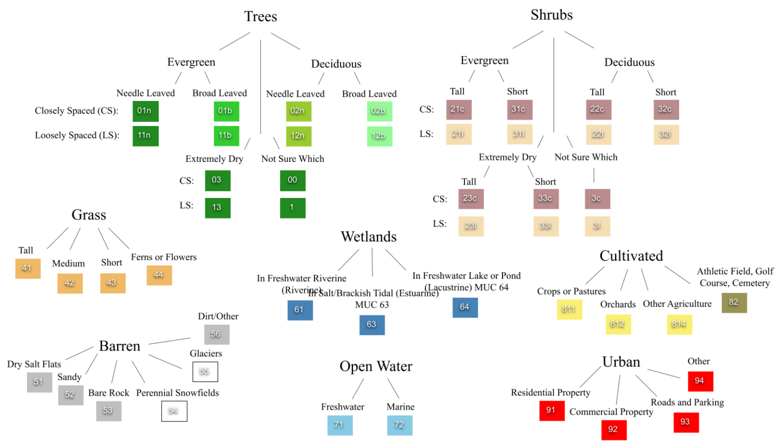

Figure 4. GLOBE Observer Land Cover categories.

more as they see fit at this stage. Furthermore, each of these sliders are colored to match their MUC code color, with the code name displayed in the top left-hand corner. Each of the sliders have the specific land cover feature name on the top and the general land cover class written out below. The minimum value for the sliders are 10% while the maximum value is 100%, with the percentage value written out next to the bar that users may move to increase or decrease the percent. The percentages chosen by the user are also allowed to amount to less than or more than 100%. When the user clicks the Next button to label the next image, the land cover features, and their percentages from the previous image transfer, thus the user is able to adjust accordingly.

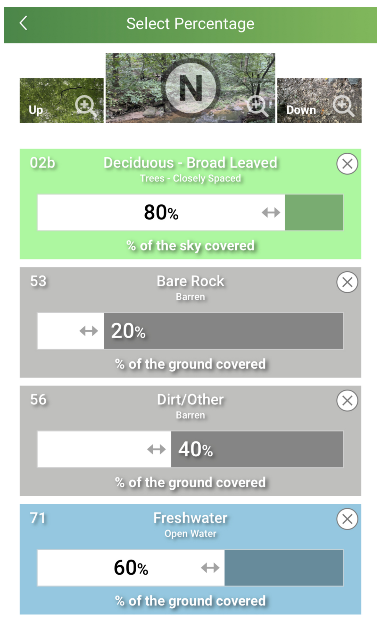

Figure 5. Percentage selection screen.

After each of the images have been classified, as mentioned prior, the user is presented with all of the information they have provided for this land cover classification. At the top of this page, there is a header called Overall Land Cover, which provides the name of the most common land cover feature described in the image classifications, along with its MUC code. Once the user has reviewed each section: Time and Date, Location, North (which includes a preview of the image and its percent classifications for each of the other directions as well), East, South, West, Field Notes, Up, and Down, they may click the Finish button to send their observation to the GLOBE database. They may also choose to continue making observations and review or send their observations at a later time using the Review/Send Observations button.

2.3 Labeling Images on Zooniverse

All 5,896 images from our team's shared Google Drive folder were uploaded into Zooniverse as a subject set. Before uploading, several other test subject sets were created to ensure that the images would be accepted without any complications. Size restrictions (1,000 KB per image) were an issue faced by the team during these tests. Pictures taken on modern smartphone cameras often have file sizes of 1-2 MB per image. This size consideration was one of the many factors that drove the decision to resize the images to 512 x 512 pixels. It took approximately 5 hours for the images to upload.

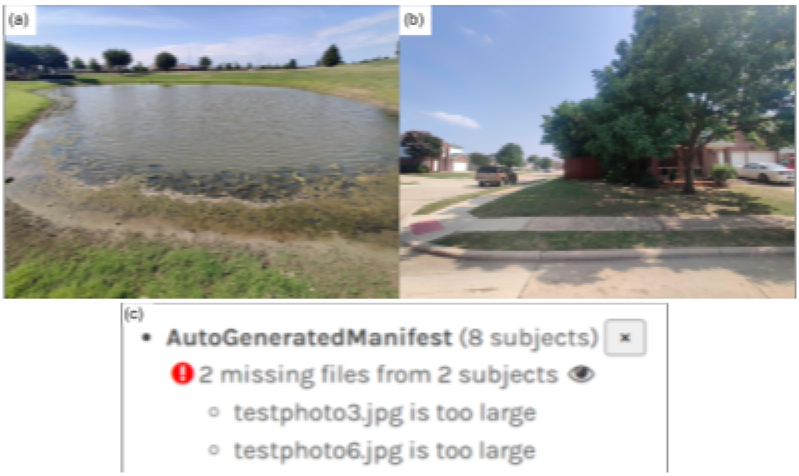

Figure 6. (a) testphoto6.jpg; (b) testphoto3.jpg; (c) Response from Zooniverse's Subject Set creation page after attempting to upload two images that exceed the size limit.

Once Zooniverse had finished uploading the images, and the subject set was complete, we began to work on the workflow. Designing the workflow was a simple process, requiring less than an hour to complete. It should be noted that our research team was given Collaborator access to mentor Peder Nelson's 'Land Cover elements' Zooniverse to use as a guide. To construct the workflow, we entered a Workflow Title, attached our subject set, and configured our workflow options.

For this project, we set the threshold for Subject retirement to a Classification count of 1, meaning that an image would only be labeled once before it was retired and taken out of circulation. Additionally, we turned on the Pan and Zoom setting, so users were allowed to take a better look at the subject they were classifying. Lastly, we hid the CLASSIFICATION SUMMARIES. This way users could move onto the next subject instantly upon completion of the current one, instead of being forced to click through a recap of their classifications.

Afterward, we created a task for the workflow. To begin, we designated the task as a Question and titled it with the Main Text text box. Upon entering the three feature categories we wanted volunteers to label (sky, land, and water), we turned on the Allow Multiple option so users could select all the features they saw in each image. Following a brief test of the task, we activated it within the workflow.

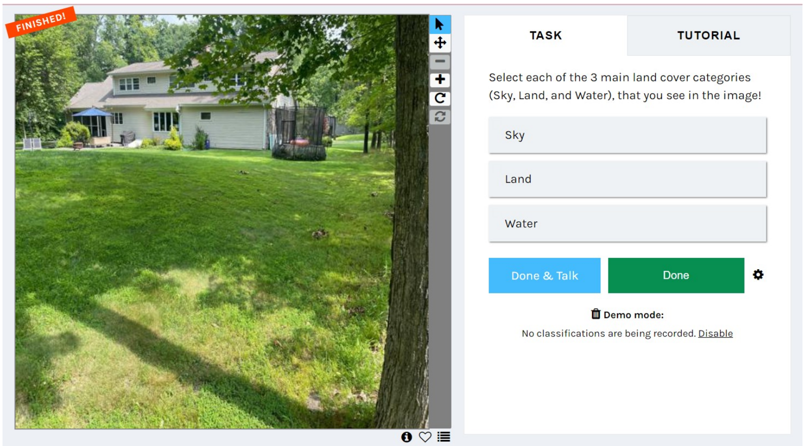

Figure 7. Labeling interface of the Zooniverse project.

Next, we constructed a tutorial. The tutorial contained photos that might confuse the volunteers, the correct labels for each photo, and why each photo should be labeled that way. We then associated that with the workflow as well.

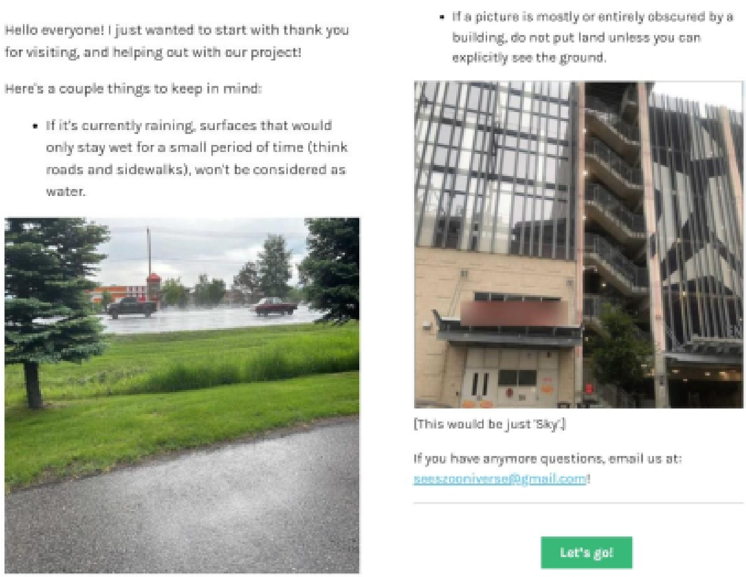

Figure 8. The Zooniverse project's tutorial.

The Zooniverse project was shared with a URL link and it took less than 4 hours to label all 5,896 images. We received a total of 5,996 classifications, due to unknown bugs in the Zooniverse software. These duplicates were later filtered out with another Python algorithm, ensuring there was only one classification per image and a total of 5,896 images.

Figure 9. Analyze Ground Observed Land Cover Photos! Workflow Statistics.

2.4 Developing CNNs to Classify Sky, Land, or Water

To achieve our desired output of identifying the presence of sky, land, or water under time constraints, we opted against using image segmentation. Although Zooniverse allows for volunteers to annotate using a polygon drawing tool, we found that the format of the export was difficult to manipulate, and it required additional time that the team did not have. As a result, we chose the ResNet-18 as our CNN architecture for its accuracy, efficiency, and straightforward implementation.

2.5 Using CNNs to Calculate Percentages

We tackled the task of suggesting percent classification for seven classes by beginning with three classes: sky, land, and water, for which we already had Zooniverse labels. To predict the percentage of each land cover category, each image was split into a grid with edge size and overlapping squares as tunable parameters. Then, the three trained CNNs to predict sky, land, and water in were used in each grid square. After the model determined whether each square contained sky land or water, we calculated the total area of a land cover category in the photo using the equation seen below.

$$Area_i = \#_{squares-classified-i} * Area_{square}$$

$$i = class : sky, land, water$$

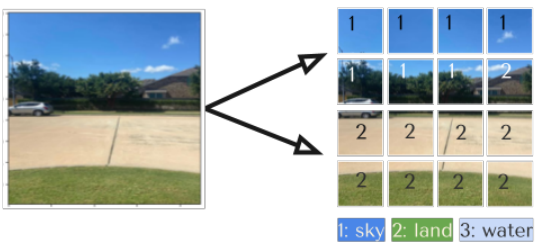

Figure 10. Diagram of the process used to derive feature category percentages.

3. Results

3.1 Performance in Detecting Sky, Land, or Water

3.1.1 Sky Detection Model

The sky detection model appeared to be one of the better-performing models, with an overall accuracy of 88%. The sky detection model performed with high metrics all across the board. Both the Area Under the Receiver Operating Characteristics (AUROC) and the Area Under the Precision-Recall curve (AUPRC) were relatively high, with accuracy percentages sitting at 94% and 98%, respectively. This appears to be the result of the abundance of sky images in our data set. Thus, our model was well-trained to detect the presence of sky, as is evident in Figure 9. This was the result of our adequate balance, as well as our satisfactory weighting and training.

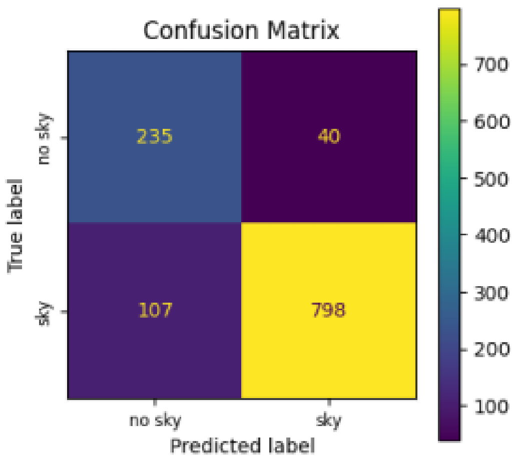

Figure 11. Confusion matrix for sky detection model.

3.1.2 Land Detection Model

Due to the fact that there were few images with no land in them, the AUROC accuracy was quite low at approximately 66%. However, our AUPRC was extremely high at about 99%. Our overall accuracy for the land detection model was 86%. This model could still be of great importance and use in the field, as there are many images taken that do not have land.

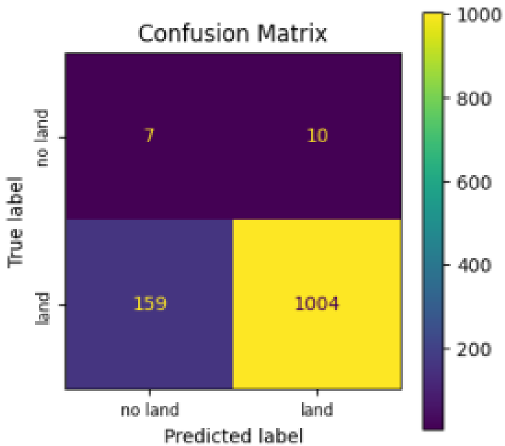

Figure 12. Confusion matrix for land detection model.

3.1.3 Water Detection Model

At approximately 94%, the water detection model had a high AUROC accuracy. However, the AUPRC accuracy was lower at about 73%, which can be attributed to the plus imbalance in the positive class. However, there was still a 0.80 true positive rate (TPR). This suggests that the model can accurately identify water most of the time, even though the AUPRC was low.

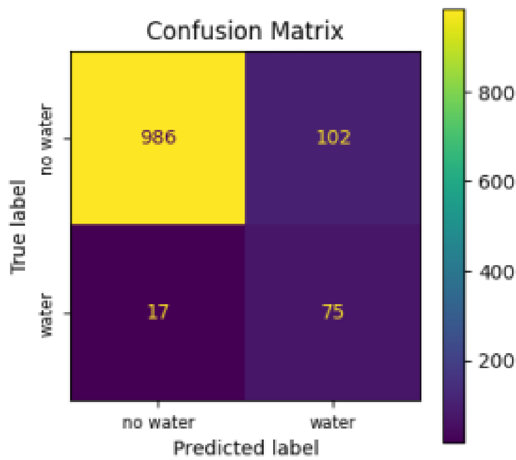

Figure 13. Confusion matrix for water detection model.

3.2 Performance in Providing Percentages for Sky, Land, or Water

3.2.1 Percentage Model

The average accuracy for the percentage model is 85%. Although the model is not perfect, it could improve land cover classifications when deployed alongside humans. Additionally, using this method to determine the percentage of a feature requires little else than Zooniverse "present"/"not present" labels, allowing the methodology to scale efficiently and easily to additional features.

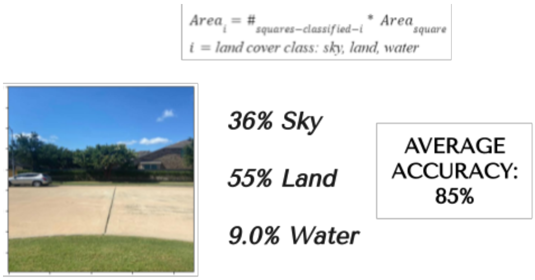

Figure 14. Equation for calculating image percentages, and an example image.

4. Discussion

4.1 Implications

A citizen science approach to training a machine learning model is explored and proven to be effective in this study. Additionally, our research explores a possible approach to developing a 'feature suggestion model' to aid citizens. Nonetheless, a model that can assess whether or not sky, land, or water is present in a ground photo of land cover can provide several benefits.

While previous research has focused on developing a model to aggregate labels across several GLOBE Observer Land Cover directional images Manzanarez et al., 2022, this model can specifically identify whether or not certain feature categories are contained in an image. The simpler output of this model targets untrained volunteers. One potential implementation would be providing users with suggestions regarding the presence of land cover features and their relative percentages in citizen science apps like GLOBE Observer.

Additionally, the research methods and strategies explored during our research demonstrate a quick and inexpensive way to gather and label large amounts of data. By allowing volunteers to gather data in their communities and label them online, citizens are provided the benefits of engaging in meaningful scientific research.

4.2 Future Work

In the future, we plan to improve our model to more accurately predict the presence of sky, water, and land categories in GLOBE Observer images. Additionally, the model will be expanded to be able to detect and classify the presence of all of the land cover categories in the seven classes recognized in the GLOBE Observer Land Cover tool for more specific feature identification. This model will also output the percentages that each feature composes in the image (sky cover for trees and ground cover for the other features). To combat data imbalances, we will adopt more consistent labeling practices and implement more comprehensive data augmentation practices. The updated model will be implemented on a mobile platform for ease of accessibility—a companion app to GLOBE Observer.

Code

The code used in this project can be accessed here: https://github.com/Centrattic/SEESZooniverse

Competing Interests

The authors declare that they have no competing interests.

IVSS Badges

4.3 I Am A Collaborator

The team members in this project are: Oseremen Ojiefoh, Roayba Adhi, Riya Tyagi, Haitham Ahmad, Lavanya Gnanakumar, and Kavya Ram.

Oseremen Ojiefoh came up with the team's research idea, explored preliminary model approaches, and designed the team's Zooniverse project. In addition to this, he contributed to the initial drafts of the 'Abstract', 'Introduction', 'Developing CNNs to Classify Sky, Land, or Water', 'Using CNNs to Calculate Percentages' portions of the paper, and wrote and developed figures for the 'Labeling Images on Zooniverse', and 'Implications' portions of the paper. Oseremen also worked on the presentation, and assisted with the proofreading and editing of the final paper.

Roayba Adhi developed the script to download images from the GLOBE API and tested the Zooniverse software. She also worked on the poster and presentation accompanying this project. In addition to this, she contributed to the initial drafts of the 'Abstract', 'Introduction', 'Performance in Detecting Sky, Land, or Water' and 'Future Work' portions of the paper, and wrote and developed figures for the 'Labeling Images on GLOBE Observer' portion of the paper. Roayba also assisted with the proofreading and editing of the final paper.

Riya Tyagi organized the images onto a pandas database, assigned labels from the Zooniverse exports, and prepared the data set for training. She also designed, trained, tested, and improved the feature detection models, as well as designed and wrote the algorithm to determine percentage land cover. In addition to this, she developed the figures for, and finalized the 'Developing CNNs to Classify Sky, Land, or Water', 'Using CNNs to Calculate Percentages' and the 'Results' portions of the paper. Riya worked on the presentation as well, and assisted with the proofreading and editing of the final paper.

Haitham Ahmad worked on the poster and presentation, facilitated the team's video voice-over, and wrote the initial drafts of the 'NASA SEES 2023 Data Set' portion of the paper. Haitham also assisted with the proofreading and editing of the final paper.

Lavanya Gnanakumar worked on the team's survey, and wrote the initial drafts of the 'Future Work' portion of the paper.

Kayva Ram also worked on the team's survey, and wrote the initial drafts of the 'Performance in Detecting Sky, Land, or Water' portion of the paper.

Everyone contributed to the revising and editing of our research paper, presentation, and poster.

Without our collaboration, we would have been forced to cut corners or decrease the scope of our research.

4.4 I Make An Impact

Our project was inspired by the difficulties faced by other NASA SEES Interns whilst classifying on GLOBE Observer. By taking the first step in developing a land cover feature suggestion model, and exploring the efficacy of using volunteers to label its training data, we hope to reduce uncertainty in classifying land cover and improve research outcomes for citizen scientists in our communities and around the world. Our efforts to automate GLOBE Observer classifications could aid in tracking changes in land cover through more efficient land cover monitoring, and assist researchers in analyzing its effects.

4.5 I Am A STEM Professional

To complete our project, we consulted with professional mentor Peder Nelson. He helped us leverage the power of citizen science by providing us with his Zooniverse project, 'Land Cover elements' as a guide to create our own Zooniverse workflow. Hours of discussion with Peder Nelson gave us new routes for exploration, enabling us to expand the work. Additionally, we spoke with peer mentor, Aswin Surya, for guidance on choosing a machine learning approach and model.

4.6 I Am A Engineer

To help volunteers classify their photos efficiently and with higher accuracy, we considered several options. Proposing a more comprehensive tutorial would only benefit those willing to sit through them, and developing more descriptive field guides might slow observers down and detract from the 'on-the-go' premise of the app. As a result, our team reasoned that engineering a machine learning model was the most effective way to approach such an individualized problem. By curating our data set, training our model, and optimizing our parameters to improve predictions, we used engineering principles to identify a problem, brainstorm possible solutions, and develop a working prototype.

4.7 I Am A Data Scientist

Our research includes the retrieval, manipulation, and analysis of GLOBE Observer data. From writing scripts to accessing GLOBE's API, to organizing our photos in a pandas database, to assigning binary classifications based on Zooniverse labels, our project makes heavy use of volunteer-provided and volunteer-labeled data. However, one limitation to our data set would be the fact that all photos were taken by the SEES23 GLOBE Observer Team. As a team consisting of interns and a couple of mentors, our photos and classifications are likely of higher quality than the general public. This means all of the photos were submitted during the duration of the internship, which took place entirely over the summer. This resulted in our data set lacking photos captured in other seasons. Moreover, there were data imbalances among the three land cover classes, with most of the images containing land and sky, and few with water. Although measures were taken to combat the imbalances and produce leveled results, our dataset would improve with the presence of more images for each individual class.

Acknowledgement

The authors would like to thank the NASA SEES Program and their mentors for their guidance throughout their research: Dr. Russanne Low, Peder Nelson, Cassie Soeffing, Andrew Clark, Dr. Erika Podest, AJ Caesar, and Aswin Surya.

The authors would like to acknowledge the support of the 2023 Earth System Explorers (ESE) Team, NASA Science Mentors, and ESE peer mentors. NASA STEM Enhancement in the Earth Sciences (SEES) Virtual High School Internship program. The NASA Earth Science Education Collaborative leads Earth Explorers through an award to the Institute for Global Environmental Strategies, Arlington, VA (NASA Award NNX6AE28A).

The SEES High School Summer Intern Program is led by the Texas Space Grant Consortium at the University of Texas at Austin (NASA Award NNX16AB89A0).

These data were obtained from NASA and the GLOBE Program and are freely available for use in research, publications, and commercial applications.

References

Alshari, E. A., Abdulkareem, M. B., & Gawali, B. W. (2023). Classification of land use/land cover using artificial intelligence (ann-rf). *Frontiers in artificial intelligence*, *5*. https://doi.org/10.3389/frai.2022.964279

Frigerio, D., Richter, A., Per, E., Prūse, B., & Vohland, K. (2021, January 1). *Citizen science in the natural sciences*. https://doi.org/10.1007/978-3-030-58278-4_5

Kohl, H. A., Nelson, P. O., Pring, J., Weaver, K. L. K., Wiley, D. M., Danielson, A. B., Cooper, R. M., Mortimer, H., Overoye, D., Burdick, A., Taylor, S., Haley, M., Haley, S., Lange, J., & Lindblad, M. E. (2021). Globe observer and the go on a trail data challenge: A citizen science approach to generating a global land cover land use reference dataset. *Frontiers in Climate*, *3*. https://doi.org/10.3389/fclim.2021.620497

Manzanarez, S., Manian, V., & Santos, M. (2022). Land use land cover labeling of globe images using a deep learning fusion model. *Sensors*, *22*(18), 6895. https://doi.org/10.3390/s22186895

The GLOBE Program. (n.d.). *Get the app - globe observer*. https://observer.globe.gov/about/get-the-app

Yuh, Y. G., Tracz, W., Matthews, H. D., & Turner, S. (2023). Application of machine learning approaches for land cover monitoring in northern cameroon. *Ecological Informatics*, *74*, 101955. https://doi.org/10.1016/j.ecoinf.2022.101955

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2018). Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(6), 1452–1464. https://doi.org/10.1109/tpami.2017.2723009