Uganda Sanitation for Health Activity (USHA)
Artificial Intelligence and Machine Learning as an alternative to surveys to Determine Sanitation Service Levels

Context

Rural sanitation programs have historically struggled to generate high quality data needed to track implementation effectiveness and associated outcomes, yet valid data is required to accurately track progress, inform planning, and investment decisions. While geo-referencing coupled with observational data collected pre- and post-intervention can be used to document changes in a household's sanitation facility and related behaviours, the process of accurately classifying a sanitation facility per international definitions is prone to error during data collection. Working with large data sets typical of rural sanitation programs hinder the process of conducting data quality assessment checks. Traditionally, machine learning and image classification have been the domain of trained data scientists and experts. There are few documented applications of ML tools in the WASH sector. USHA seeks to compare outputs of sanitation service ladder classifications by enumerator classification and ML Image classifications.

Overview

In early 2019, the USAID Uganda Sanitation for Health Activity (USHA) applied the Ministry of Health's National Sanitation Market Guidelines for Basic Sanitation (NSMG) to design a novel intervention aimed at increasing household investment in basic sanitation in two customer segments. Coined the Market Based Sanitation Implementation Approach (MBSIA), the intervention has been implemented in 13 districts in the Central and Eastern regions, targeting 219,843 households within 1,958 villages. A Community Led Total Sanitation (CLTS) with quality approach was also implemented in seven northern districts, targeting over 49,380 households in 878 villages. Both the MBSIA and CLTS programs used pre- and post-intervention surveys to document household sanitation service classifications with GPS and pictorial evidence.

In early 2021, USHA, with support from Tetra Tech's Technology for Development team explored the use of artificial intelligence and machine learning (AI/ML) applications to categorize images of newly constructed or upgraded toilets to determine if the toilets met the WHO/UNICEF/Joint Monitoring Program (JMP) minimum standards for household sanitation services. The activity was interested in three classifications: type of superstructure, washability of the interface, presence/absence of a door. Over 270,000 images of latrines (e.g. interface and superstructure) were collected during implementation yet the team did not have the capacity to process and analyze the data to confirm the status of the toilets. To address this, USHA applied a machine learning model as the processing capabilities of AI/ML tools significantly reduced the time required to classify and analyze large amounts of data captured in the toilet images collected by USHA.

Methodology & Results

ML & image classification analysis utilizes Lobe.ai, TensorFlow, and Python to complete the full classification and evaluation of the images collected via the USHA MBSIA and CLTS+ baseline and endline surveys. Lobe.ai is a free AI image analysis software developed by Microsoft and can be used to generate image classification models. TensorFlow is a neural network platform that develops and

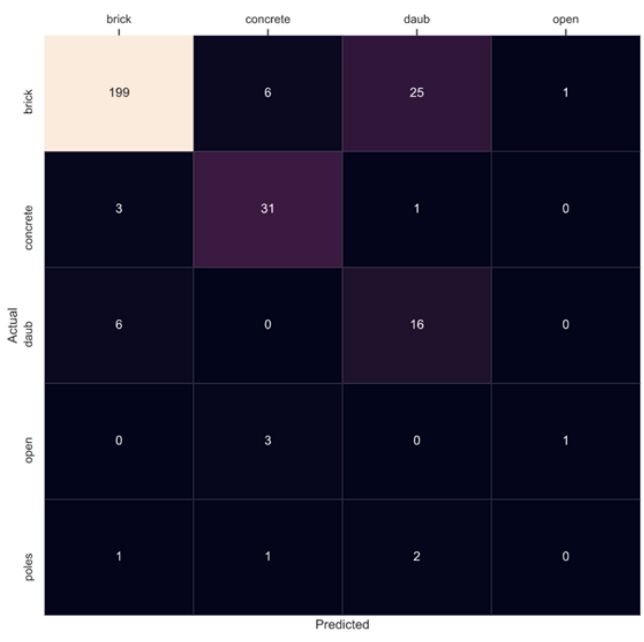
deploys largescale machine learning models using classification models. Python is a computer language that instructs TensorFlow how to execute the model. The data analysis and model development process begins with labelling of images using a convenient labelling system or by sorting them into digital folders. Each label corresponds to a category of image, i.e. all brick superstructures get the "brick" label and all wattle and daub structures get a "daub" label. Second, Lobe.ai is utilized to analyse the image to create a model that can be used to predict labels for new images. As part of the data analysis, Lobe.ai iterates the model in real time by comparing the results of new images analysed against images that have already labelled. This allows the user to assess the accuracy of the model and determine if additional data is required to improve the model before the model is applied to the full dataset. Once the Lobe.ai models are deemed suitable, they are into TensorFlow where large sets of images can be analysed in accordance with the Python script.

There are two methodologies, one for baseline and one for the endline results. While the baseline model for floor classification was accurate it focused largely on the material. It was determined that the endline model should focus on washability of the floor. Two new models were created for the endline survey results that focused whether the floor is considered washable or unwashable. The following reviews the methodology used for the baseline and endline machine learning models.

Baseline Model Methodology

We first tested Lobe with structure classifications. A sample of images were fed into the system with structure classification labels for different materials such as brick, mud-brick, concrete, daub, and open (no structure present). This yielded poor labelling performance of <60% and thus required additional samples to increased the rated accuracy to 90-95%. Once the desired accuracy rate was achieved, we evaluated the feasibility of utilizing this model to analyse the 8,000 images collected to date.

Figure 1:



We developed two Python scripts that linked the survey platform Ona.io where all latrine images were stored to the TensorFlow model. The first Python script (*image_downloader.py*) downloads and resizes the images from Ona.io and saves them locally. A second script loaded the model produced by Lobe, ran it on each image, and produced a CSV with the classified labels (*run_model.py*). With this proof of concept complete, we were able to improve the model accuracy and scaling to run the model on more images.

Several hundred more images were labelled for each of three classification tasks: superstructure, floor material, and door presence. Lobe reported accuracy of 92-96% for each of these. The script was then modified to allow a user to choose which model to run, and a directory of images to run it on.

At this point, we also ran the models on several hundred additional images to test our script at scale and have a larger set to validate.

We did a second pass at accuracy evaluation with the additional data that was acquired. Because Lobe does validation testing on the same dataset that it used for training, there is an opportunity to introduce overtraining, a fundamental challenge in AI/ML image classification, which is the tendency for an ML model to become very good at classifying its training data, at the expense of "real" data. Unfortunately, this proved to be the case. By hand labelling three hundred new images in a CSV file and comparing these labels to Lobe classifications, we found accuracy of 85%, compared to 95% self-reported by Lobe. This is still quite good for an image model (and about as good as the best models 5 years ago), but less than ideal for our purposes.

A *confusion matrix (Figure 1)* identified most discrepancies occurring between regular red clay bricks and mud bricks. As a result, we experimented in merging these two classifications, which improved accuracy to 89%. However, we still observed some problematic classification patterns for other combinations which can likely only be improved with additional training data. The methodology for the endline survey classification describes the adjustments made to the floor model.

Endline Model Methodology

As mentioned above, the primary goal for the endline model was to improve the accuracy of the baseline model for the "interface" or floor model. The most important aspect of the floor structure is it is considered washable. The material of the toilet "floor" or interface is one of two key defining features used globally to describe the level of household sanitation services (the other being if the facility is shared). Toilets with washable interface materials (concrete, tiles, plastic, etc.) are considered to be improved facilities.

Two models were created (one targeting the Northern Cluster and the second targeting the Central East (CE) and Central West (CW) images.

Northern Cluster Model

In the NC, the categorization of floor types is important and according to JMP/WHO/UNICEF internationally accepted definitions "slabs covered with a smooth layer of mortar, clay, or mud should also be counted as improved."[1] In the NC, few interfaces are built with concrete or other washable materials but the washable designated can include smooth clay which is commonly found in the NC. In CE and CW latrines with smooth clay or mud floors would not be classified as improved but this is a common floor type in NC where building practices are different and poverty levels are higher. This is an important nuance that needed to be captured in the revision of the original models for the endline classification.

It was determined that due to these regional differences, two floor models would be developed and those models would be trained differently based on what is appropriate for the distinct regions. The most important distinction for the NC is between washable (smooth clay) and unwashable (dirt) latrine

[1] Guidance for monitoring safely managed on-site sanitation (SMOSS) Draft prepared for Phase 2 pilots August 2022

floor materials. Figure 2 below shows the classifications for each version of the NC model. Multiple iterations of the model were created due to misclassification of the various labels of the NC training images. For the final NC model (Version 4) the model had difficulty predicting images that have covers over the latrine hole on smooth surfaces. The model predicts those as "smooth" when they should be "washable." We found that adding more images to the model made a significant difference in the accuracy for Version 4. With accuracy of 90%, the most accurate training model was Version 4 which utilized 3 label types: smooth, washable, and unwashable. Once finalized, the TensorFlow model was exported and run in the Python script to be used across all images in the Northern Cluster Endline Survey CSV.

Northern Cluster Model Results

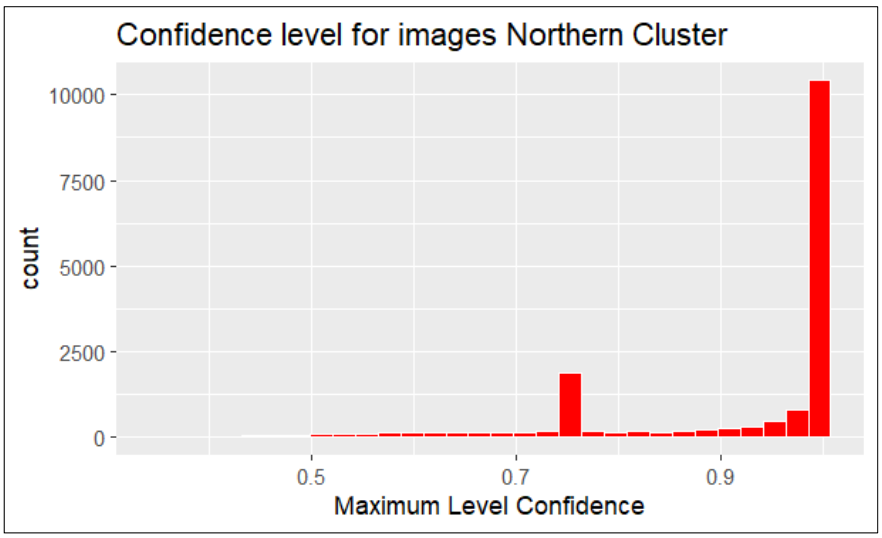
The final model for the Northern Cluster had a high 90% accuracy (see Figure 2). The most effective way to increase the accuracy of this model was to decrease the number of labels and also use of clear/high resolution images that show more distinction between washable, unwashable and smooth classifications.

Figure 2: Northern Cluster Model Results

Model Version: Northern Cluster	**Accuracy of Lobe.ai Model**	**Labels**
Version 1 (1449 images in training model)	42%	• Cement screeding • Concrete slab • Dirt floor (Unwashable) • Pit only (No superstructure) • SanPlat • Sato with washable floor • Smooth unwashable floor
Version 2 (1582 images in training model)	88%	• Unwashable • Washable
Version 3 (890 images in training model)	65%	• Dirt Unwashable • Smooth Clay • Washable
Version 4 (1523 images in training model)	90%	• Smooth • Washable • Unwashable

After running the TensorFlow model on the Endline images, the confidence level is one of the outputs. This confidence level is the main accuracy indicator and shows how confident the model is in the classifications by image. As seen in Figure 3, the confidence level for the majority of the images is close to 1 (on a scale of 0-1) which shows the model is highly confident/accurate for the Northern Cluster.

Figure 3: Northern Cluster: Confidence Level of Model by Image Count



The accuracy of the model by classification shows that in the endline survey 29% of the total images are now categorized as washable with a High level of confidence. Among the latrines classified as washable, 81% have a high level of accuracy for the classification. Figure 4 shows the accuracy of the Smooth, Unwashable, and Washable classifications by using different levels of confidence levels: High is 95% confidence, medium is 85-95%, low is 50-85%, and very low is below 50%. The model results are quite accurate with most of the images falling in the High and Medium categories.

Figure 4: Northern Cluster: Classification and Accuracy of Model

Model Classification	Accuracy	Count of Images (17,256 total images)	Percent of Total
Smooth	High	2970	17%
Unwashable	High	3532	20%
Washable	High	5001	29%
Smooth	Med	367	2%
Unwashable	Med	295	2%
Washable	Med	454	3%
Smooth	Low	554	3%
Unwashable	Low	2502	14%
Washable	Low	692	4%
Smooth	Very Low	56	~0%
Unwashable	Very Low	72	~0%
Washable	Very Low	38	~0%

Central East/Central West Model

The Central East/Central West (CE/CW) interface model was based on the original baseline model. Figure 3 below illustrates the labels used in the predictive Lobe.ai model. Version 1 of this model had difficulty predicting cement screeding and concrete slab. Additionally, there was redundancy and overlap in the images which caused the model to incorrectly predict the "washable" labels. Version 2 was much improved once only 2 labels were utilized. The accuracy increased to 95% and this model was used in the Python script against the Endline Survey data for the CE and CW regions. The final model for the Central East/Central West Clusters had a high 95% accuracy (see Figure 5). As in the Northern Cluster model, the most effective way to increase the accuracy of this model was to decrease the number of labels to avoid confusion in the model as many of the previous labels were not mutually exclusive. This caused confusion in the model and the most important aspect of the floor model is whether or not it is washable or not washable. This is reflected in the final version.

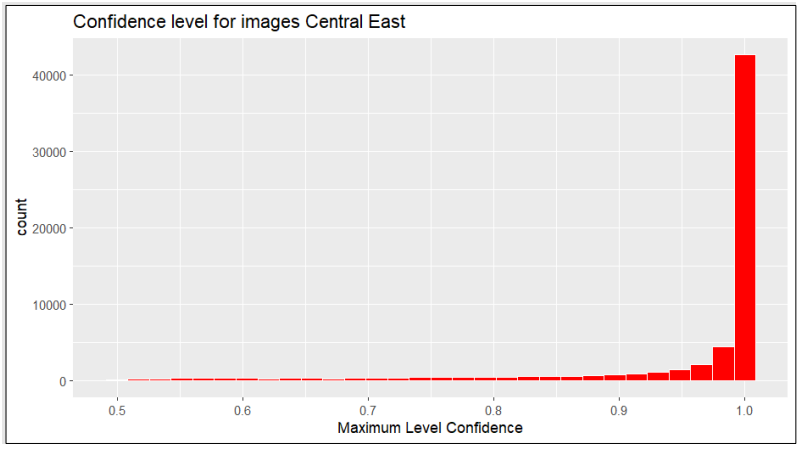
Figure 5: Central East/Central West Model

Model Version: Central East/Central West	**Accuracy of Lobe.ai Model**	**Labels**
Version 1 (698 images)	82%	• Cement screeding • Concrete slab • Dirt floor (Unwashable) • Pit only (No superstructure) • SanPlat • Sato with washable floor • Tile
Version 2 (1582 images)	95%	• Unwashable • Washable

Central East Model Results

The confidence level for the majority of the images in the Central East endline survey is close to 1 (on a scale of 0-1) which shows the model is highly confident/accurate for this region (Figure 6).

Figure 6: Central East: Confidence Level of Model by Image Count



The accuracy of the model by classification shows that in the endline survey 48% of the total images are now categorized as washable with a High level of confidence. Among the latrines classified as washable, 91% were classified with a high level of accuracy. Figure 7 shows the accuracy of the Unwashable and Washable classifications by using the following levels of confidence levels: High is 95% confidence, medium is 85-95%, low is 50-85%, and very low is below 50%. The model results are very accurate with most of the images having a High confidence level.

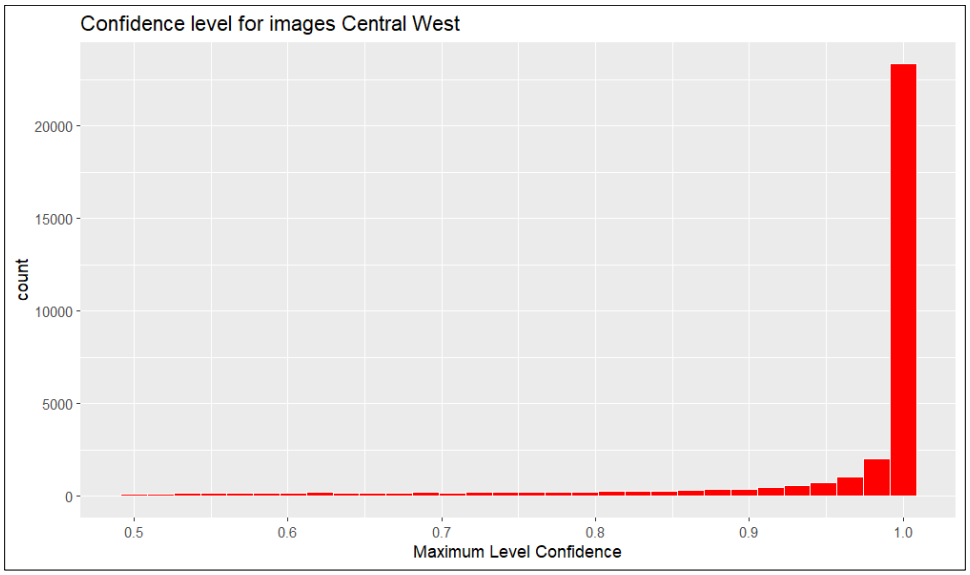
Figure 7: Central East: Classification and Accuracy of Model

Model Classification	Accuracy	Count of Images (62,913 total images)	Percent of Total
Unwashable	High	19758	31%
Washable	High	30159	48%
Unwashable	Med	2197	3%
Washable	Med	3002	5%
Unwashable	Low	3659	6%
Washable	Low	4138	7%

Central West Model Results

The confidence level for the majority of the images in the Central West endline survey is close to 1 (on a scale of 0-1) which shows the model is highly confident/accurate for this region (Figure 8).

Figure 8: Central West: Confidence Level of Model by Image Count



The accuracy of the model by classification shows that in the endline survey 55% of the total images are

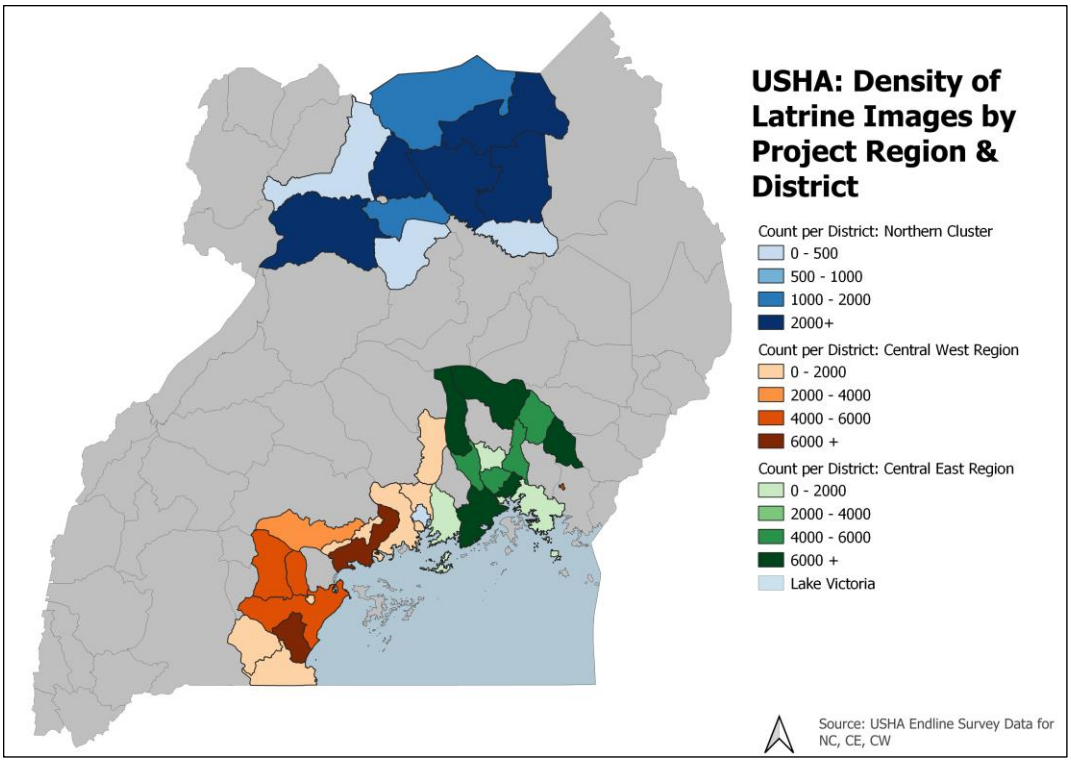
now categorized at washable with a High level of confidence. Figure 9 shows the accuracy of the Unwashable and Washable classifications for Central West region by using the following levels of confidence levels: High is 95% confidence, medium is 85-95%, low is 50-85%, and very low is below 50%. The model results for Central West are the most accurate among the three regions with most of the images having a High confidence level.

Figure 9: Central West: Classification and Accuracy of Model

Model Classification	Accuracy	Count of Images (32,908 total images)	Percent of Total
Unwashable	High	8687	26%
Washable	High	18069	55%
Unwashable	Med	948	3%
Washable	Med	1534	5%
Unwashable	Low	1685	5%
Washable	Low	1985	6%

Conclusions

Once the final TensorFlow "floor" models were exported from Lobe.ai, they were utilized in the run_model.py script for the final classification of the Endline Survey images. The map below (Map 1) shows the density of the images by district and further divided into region (NC, CE, CW). This shows not only the coverage of the USHA project but also the enormous amount of imagery data that was leveraged for this study.



Map 1: Density of Images by District and Region based on GPS points in Endline Survey

Enumerator vs. Machine Learning classification of Latrine types

Another key element of this analysis beyond the accuracy of the model in the three regions is to compare the classifications from the Enumerators toilet image selection and their field data to see how closely it matches with the model's classifications. Figure 10 shows the results of this analysis: if the two columns of the Enumerators' classification and Model Output columns were the same (I.e. both washable or both unwashable) then the id is classified as a "match". If the column classifications were different, for example if the enumerator classified a latrine as washable but the model classified the latrine as unwashable, then it was "not matched." 70% of the classifications were a match for Central East, 76% for Central West, and only 46% for Northern Cluster. The Central East and West regions have very positive results which shows that in the future, a model like the one developed for this analysis could be used in place of enumerators. For example, if beneficiaries sent an image of their latrine to a central digital location it could decrease the resources and time needed to send agents into the field to collect data. This is one of the positive outcomes of machine learning and artificial intelligence especially in hard-to-reach areas or conflict zones. With only 46% of the total being a match for the Northern Cluster could show that either the model needs to be further refined or the method of classification by the Enumerators needs to be clarified. As mentioned above, even latrines with a "smooth or clay" surface are considered washable in this zone, this distinction needs to be clarified with Enumerators to ensure consistency for classification purposes.

Figure 10: Comparison of Enumerator and Model Output Classification

Region	Not matched	Matched	Data not available	TOTAL	Percent matched of total
Central East	18214	45248	1584	65046	70%
Central West	7366	25800	620	33786	76%
Northern Cluster	7984	8014	1258	17256	46%

Another important element in determining whether a latrine is improved or unimproved is whether it is shared by more than one household. Using the Endline survey data, Figure 11 shows the breakdown of shared versus not shared totals for each of the households. Most households in this analysis do not share their latrine.

Figure 11: Shared or not shared

Region	Not Shared	Shared	TOTAL	Percent of total: NOT SHARED
Central East	56157	7305	63,462	88%
Central West	29568	3598	33,166	89%
Northern Cluster	10722	5908	16,630	64%

The second level comparison for the endline survey in each region included the following questions: Is the latrine washable or unwashable and is the latrine shared or not shared? The results of this analysis are shown in Figure 12. Some of the highest percentages in this analysis for all regions are the "Washable and Not Shared" category. This represents the improvements made under the UHSA project toward improved sanitation for beneficiaries.

Figure 12: Comparison of Enumerator and Model Output Classification Washable and Shared, Percent of Total classification that is a match between Enumerator and Model Output classification

Region	Washable and Not Shared	Washable and Shared	Unwashable and Not Shared	Unwashable and Shared
Central East	50% match	7%	35%	4%
Central West	56% match	8%	30%	3%

Northern Cluster	22% match	13%	24%	13%

Lessons Learned

- ML and Image classifications works best with clearer high-resolution images of toilet floor types
- Enumerators should ensure that drop hole area of the latrine is not covered to avoid any obstructions of the ML models while classifying images
- The use of ML & Image classification models to classify latrine types can save time and resources required to administer long surveys
- Sanitation service surveys should further classify "unwashable floor" materials by durable smooth mud/mortar or not to ensure compliance with the JMP/WHO/UNICEF standards especially in low income earning communities such as Northern Uganda.

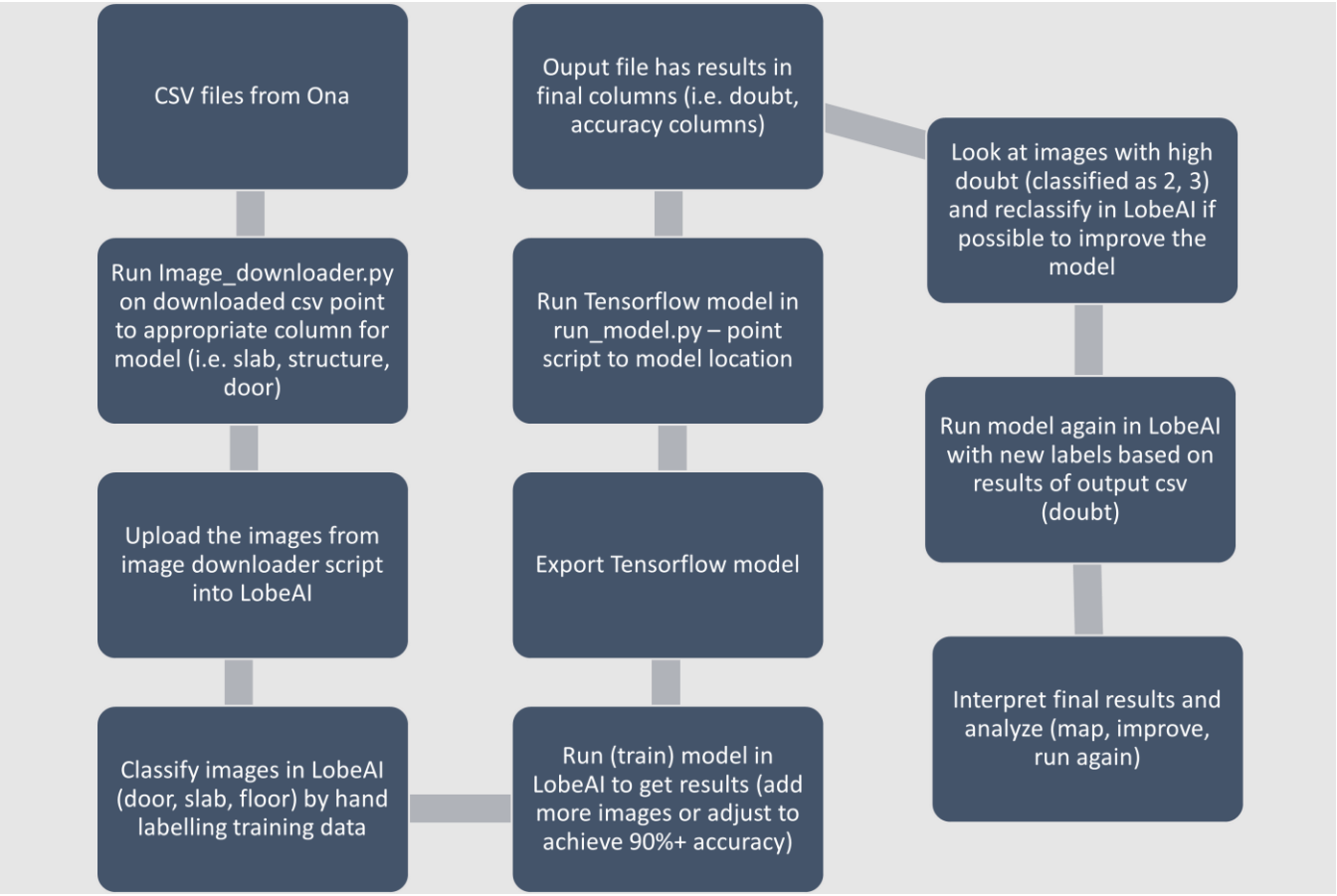
Conclusions

This analysis aimed to show how machine learning could be leveraged to help classify large-scale sanitation projects with accuracy and high levels of confidence. USHA is keen on sharing this AI/ML model with the Implementing Partners, for Water and Environment, and Ministry of Health for use and further adaption as an alternative to determining sanitation service ladders using image classification.

The TensorFlow models for the Floor, Superstructure, and Door are publicly available on the USHA Google Drive and the Python code and methodology is available on GitHub. Machine learning models require extensive adjustments to avoid things like overfitting and poor accuracy. The iterations on this model and the overall approach to the classification of latrine floors, structure, and door presence are captured in Appendix 1 (Workflow Design). Appendix 2 describes how to run the Python Scripts.

Appendices

Appendix 1: Workflow design



Appendix 2: Guidance on Using Python Scripts for AI/ML

Programming projects require consistent and structured files and folders to decrease error rates and improve efficiency when using the model and associated outputs. The following instructions are important to ensure these models run successfully.

File and Folder Structure

Create a permanent folder on your hard drive which will serve as your base folder for the model, add the following subfolders and scripts to this base folder:

1. csv - to store the CSV exports from Ona
2. images – subfolders for each region can be created under this folder; this is where the model will store downloaded images
3. models – stores the Lobe models
4. outputs – model outputs are placed here (CSV format)
5. src – extra code files
6. image_downloader.py – this script will download batches of images from Ona, and shrink/resize them for the model
7. run_model.py – this script runs on a folder of images for the selected model (door, floor, structure), leverages the TensorFlow model, and outputs classifications

Download images

1. Export a CSV file from Ona, save it in the csv folder, and make note of which column has the images you want to classify. The columns have links to the Ona survey folder. Suggest renamed the columns to single words like 'superstructure' and 'interface' just so you do not have to write out 'q13b etc etc etc'
2. Make a subfolder inside the images folder to store the downloaded images and point the Python Script to that folder for where to save the images
3. Open up command prompt. Two ways to do this:
    a. Hold down shift and right click, then choose *Open command window here*
    b. Run *Command Prompt* from the start menu and then navigate to your folder by typing *cd your/folder/path/here*
4. In the command prompt type *python image_downloader.py*
5. It will ask you for the name of the CSV file, the column with the image URLs in it from 2a above, and which subfolder you want to store in
6. It will start downloading images and saving to the subfolder. When it stops, it's done!

To run the model

1. Open a command prompt as above and type *python run_model.py*
2. It will ask you for the folder the images are in, and which model you want to run by choosing 1, 2 or 3. You can also provide the CSV file again and the column name with the URLs in that file, to be merged with the final output.
3. Once you hit enter, you'll see outputs running very quickly. Once it's stopped, it's done.
4. It will save a file called output-*ddmmYYYY-HHMM*.csv where those are the date and time stamp. This file will contain all the info to evaluate the prediction on the images given, and if the user provides the original CSV, the prediction evaluation information will be appended to that CSV and saved in the above format.