EYES OF CASSAVA

Leandro Cunha
Senior Software Engineer
Belo Horizonte, MG, Brazil
lvcunha@gmail.com

Pedro Lourenço
Data Scientist
São Paulo, SP, Brazil
pedro.gengo.lourenco@gmail.com

Vinicius Busquet
Senior Software Engineer
Rio de Janeiro, RJ, Brazil
vibusquet@gmail.com

Vinicius Granja
Computational Linguist
New York, NY, USA
vin.idem@gmail.com

August 14, 2021

ABSTRACT

Pests, disease and weed are major sources of poor yield for Cassava by 50 percent [4]. The current method used to identify the plant disease requires agriculture experts to do so. That is costly, slow, labor intense, and dependent on a limited number of experts. Our model solve this yield issue by detecting common diseases on the cassava leaves with OAK-D. We use a model called CropNet which was created by Google as a Benchmark to our model evaluation [1]. We concluded that our model has better generalized the results and outperformed Google's CropNet on a balanced data set by 3 percent and a set of very blurry images.

1 Introduction

Our project has the goal to help improve farmers' production yield of cassava in Brazil and countries of sub-Saharan Africa by building a cassava disease identifier model running on OAK-D. The idea of creating that solution with OAK-D came after participating in a Kaggle competition called Cassava Leaf Disease Classification [9]. We saw a potential to apply what we learned in that competition to OAK-D. Moreover, cassava plantation fields in Brazil are widely available.

Cassava is called the food of the 21st century by the UN. It is consumed by around 700 million people worldwide but most of the consumers are located in the poorest countries in the world [2]. Despite being originally from South America, Cassava has become a staple food in Africa. The plant is very rich in carbohydrates and survives very harsh weather and soil conditions. Cassava provides the 3rd highest carbohydrate yield among crop plants in tropics. In Africa, it provides the second highest carbohydrate yield among crops. It is so resistant that soils with pH ranging from acidic to alkaline can't stop it from growing. There is no problem if it rains too little or too much. It can survive on 50mm to 5m rainfall in a year. It does not need sunblock. It survives on equatorial temperatures. If that was not enough to believe how tough this plant is, it can survive on elevations between sea-level and 6600 feet high [**?**]. Because of its resistance and high source of carbohydrates, it provides carbohydrates to the most impoverished regions of the world. To put that in perspective, cassava trade is responsible for 46 percent of the GDP in Ghana. It is both a cash and subsistence crop in such impoverished countries. That is why it is called the food of the 21st century by the UN. However, pests, disease and weed can reduce its economic yield by 50 percent [4]. The current method used to identify the plant disease requires agriculture experts to do so. That is costly, slow, labor intense, and dependent on a limited number of experts. We want to solve this yield issue by detecting common diseases on the cassava leaves with OAK-D.

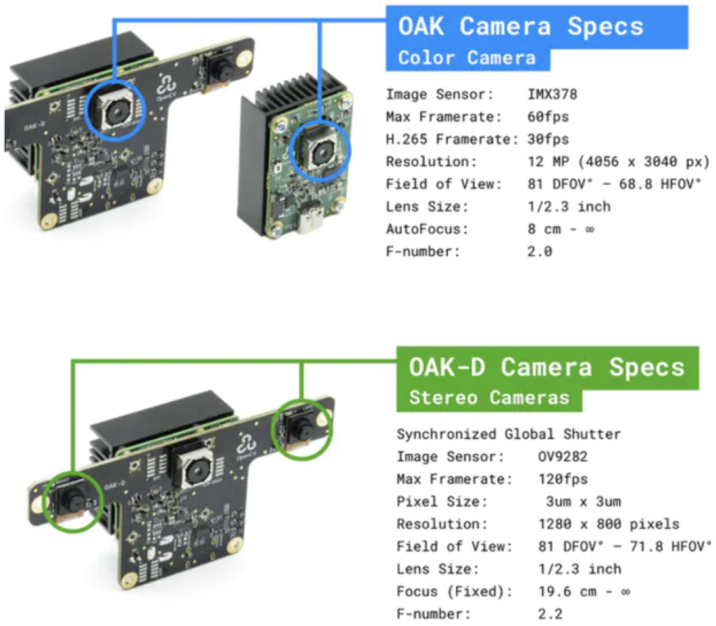

Figure 1: OAK-D camera specs [5]

2 Cassava leaf disease data set

class_name	label	frequency
Cassava Mosaic Disease (CMD)	3	3212
Cassava Green Mottle (CGM)	2	1503
Cassava Brown Streak Disease (CBSD)	1	1421
Cassava Bacterial Blight (CBB)	0	1187
Healthy	4	1166

(a) Train Set: 8490 images

class_name	label	frequency
Cassava Mosaic Disease (CMD)	3	750
Cassava Green Mottle (CGM)	2	411
Cassava Brown Streak Disease (CBSD)	1	354
Cassava Bacterial Blight (CBB)	0	334
Healthy	4	274

(b) Test Set: 2123 images

Figure 2: Train and test data sets breakdown

The data set consists of a total of 10643 labeled images of cassava leaves. There are 8490 labeled images in the training set and 2123 labeled images in the test set. The annotations are separated in 5 classes: Cassava Bacterial Blight (CBB), Cassava Brown Streak Disease (CBSD), Cassava Green Mottle (CGM), Cassava Mosaic Disease, and Healthy. The data was manually re-labeled by us because we found inconsistencies in the original labelled dataset from Kaggle. We randomly re-labeled 10643 images from the original dataset of 21,367 labeled images which was available on Kaggle and labeled by experts from the following agencies, Makerere Artificial Intelligence (AI) Lab and the National Crops Resources Research Institute (NaCRRI). Makerere Artificial Intelligence (AI) Lab is a Data Science research group

from Makerere University in Uganda. NaCRRI is a government agency responsible for agricultural research in Uganda. Bellow we explain the reasons behind the re-labelling.

2.1 Reasons for data re-labelling

We noticed a couple issues with the labeled data during the results of our modeling results. In order, to best guide us and validate our model we used the model CropNet as a benchmark to our work. CropNet was created by a Google team in partnership with Makerere Artificial Intelligence (AI) Lab and the National Crops Resources Research Institute (NaCRRI). CropNet was created on the same data but with a specific size of train, test and validation set. CropNet used 9430 labelled images split into 5656 training set, 1885 in test set and a 1889 validation set. CropNet also informed that the number of images per class were unbalanced of 72 percent of the total set with just two disease classes CMD and CBSD.

3 Diseases

3.0.1 Cassava Bacterial Blight (CBB)

The bacterial infection shows angular spots, water-soaked spots which occur due to restricted flow of nutrients by the leaves. The spots are easily seen in the lower surface of the leaf. The spots expand throughout the leaf, especially on the margins of the leaves. It later becomes yellow with brownish spots. The infected leaves do not become distorted like CMD and CGM.

3.0.2 Cassava Brown Streak Disease (CBSD)

This is the most troublesome disease among all the diseases. We will provide more details about this disease than the others due to its high yield loss to cassava plantations. The symptoms are sometimes tricky to find but devastating. If a plant is infected by it, there will be a total loss of the plant and its roots. It cannot even be used to feed animals. Symptoms of cassava brown streak disease appear as patches of yellow areas mixed with normal green colour. This phenomenon is commonly referred to as chlorosis. Chlorosis is often also associated with secondary veins; veinal chlorosis is on mature cassava leaves The yellow patches are more prominent on mature leaves than younger ones. The infected leaves do not become distorted in shape like CGM and CMD. Advanced symptoms on the leaves become an irregular yellow blotchy chlorosis that is most pronounced in the periphery (margins or edge) of lower leaves. It is these advanced symptoms that make CBSD so similar to CBB. It is at this late stage of symptoms that has caused high false negatives, to CropNet, ResNext50 and CNN models we tested.[3] The disease is spread through planting of stem cuttings from CBSD infected plants. Planting cassava cuttings from infected cassava plants is the major way CBSD spreads. Sharing and distribution of infected planting materials is responsible for rapid spread of the disease. The virus also spreads from plant to plant by white flies. Planting of susceptible varieties helps build up CBSD in the affected countries. Moreover, knives used in cutting cassava sticks into cuttings can spread CBSD to healthy planting materials when the infested knife is used on them.[4]

3.0.3 Cassava Green Mottle (CGM)

It is a fungal disease that causes distortion of the leaves with bright white or yellow small spots. At young plants the leaves might display the bright white or yellow spots without distorting the leaves. In severe cases the leaves become stunted. It is most similar to CMD; however, the coloring of the spots in a CGM is faint when compared to CMD spots which display a very vivid yellow color but in less uncommon periods an white-like color.

3.0.4 Cassava Mosaic Disease (CMD)

Cassava Mosaic Disease is characterized by severe mosaic symptoms on leaves. The affected leaves display a light green and strong yellow coloration followed by severe shake distortion of the leaves. In some cases, the spot coloration is whitened which makes it very similar to CGM in such cases of white coloration.

4 Modeling

We used two model architectures, ResNext50 and CNN, while benchmarking results were against MobileNetv3 used by CropNet. ResNext50 performed better than CNN and MobileNetv3. ResNext50 had an accuracy of 80.7 percent on a test set of 2123 images while CNN had an accuracy of 63.5 percent and MobileNetv3 had an accuracy of 77.6 percent. MobileNetv3 dropped from 88 percent accuracy with a skewed training set and validation test of 72 percent. With our

fixed skewness and relabelling of our 2123 test set led to a drop of 10.4 percent in accuracy of CropNet. We successfully surpassed CropNet with our ResNext50 architecture by 3.1 percent. Those accuracies were also based on the following augmentations. CropNet's MobileNetv3 used the following augmentations: brightness, contrast, saturation, hue, and crop. Our ResNext50 used crop, transpose, rotation, shift rotation, pixel normalization, dropout, and cutout. CNN used brightness, contrast, saturation, hue, crop, rotation, and blurriness.

5 Inefficiencies of CropNet

We focused on building and using an already built model, CropNet, as benchmark to improve our model at higher standards than the Benchmark. Our model is not only trying to surpass CropNet in terms of architecture but it is also addressing the inefficiencies our team found in the CropNet model.

5.0.1 Skewed Data

72 percent of the CropNet data only represents two illnesses, CMD and CBSD. The remaining classes, CBB, CGM and Healthy was 28 percent.

SOLUTION: We reduced the skewness from 72 to 43 percent. The remaining classes, CBB, CGM and Healthy was 57 percent. You can find those ratio results when you combine the data on Figure 2.

5.0.2 Root and Trunk images

The set has a mix of root and trunk which is very negligible.

SOLUTION: We did not make fixes to our model because it would reduce our accuracy when comparing to the CropNet model for not learning how to classify root and trunk. Moreover, all root diseases were only labeled CBSD. It is a mistake that we maintained in order to not lead to bias because we benchmark against CropNet test and validation set that had such a mistakes.

5.0.3 Significant data mislabeling

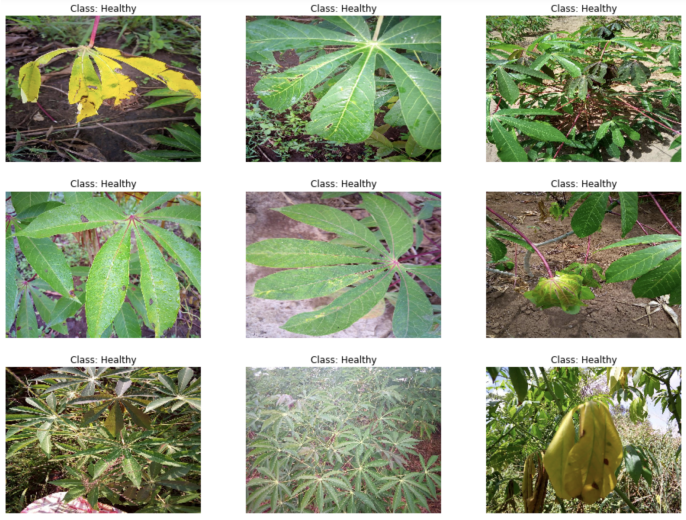

Figure 3: Miss-labeled healthy images

There are significant mislabeling of data specially in the Healthy sample of the data. How we found that out even though we are not specialists like Google and the government agencies from Uganda? We first noticed a high number of mislabeling when looking through the images manually as it can be seen on Figure 3. There were 2577 images labeled as Healthy in the Kaggle data. We decided to relabel the data and found out there was actually only 1440 Healthy leaves. The rest were leaves with diseases. We fully proved the mislabeling when we ran a confusion matrix on the results we got from running predictions on our test set with CropNet which can be seen on Figure 6. It visible the excessive false negatives on Figure 4 for the Healthy predictions column of CropNet for our 2123 test set. SOLUTION: It is because of

that large mislabeling mistake that we decided to relabel at least 10643 images from the 21,367 data set from Kaggle. Our model showed better generalization of training set when we made the relabelling of all the Healthy and disease ones, except for CMD. We only re-labeled 3962 CMD images out of 13158 which corresponded to approximately 30 percent of Kaggle's skewed data set. As a result, CropNet accuracy dropped from 88 percent, from their validation data set, to 77 percent against our data validation set whose mislabeling was fixed and CMD's data skewness was reduced by 70 percent.

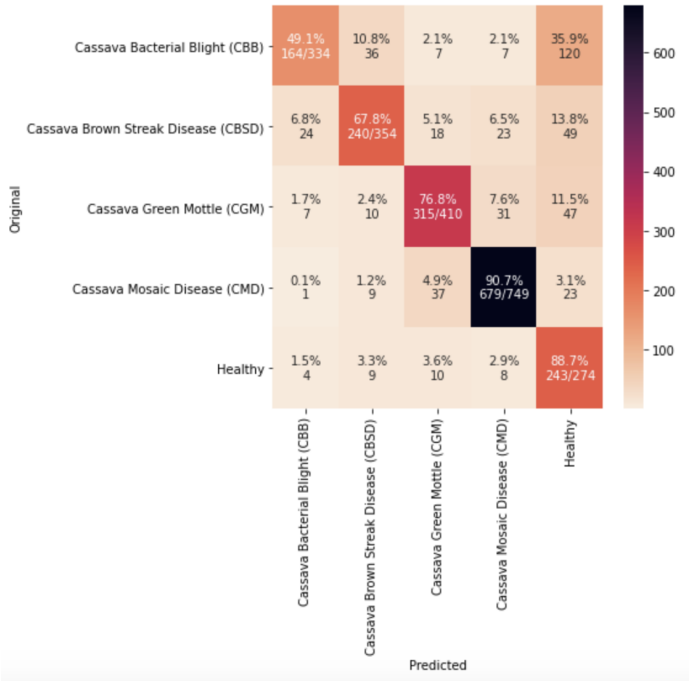

Figure 4: CropNet confusion matrix from our test set of 2123 images

5.0.4 Duplicated Images

There are duplicate images in the data set that are also have different labeling. By using the Hamming Distance model for pixel similarities, we were able to find 21 duplicates. In some cases, the duplicates shared different labels which usually leads to bias and can throw off the model when two or more of the same images are labeled differently as you can see on the first and second images of leaves on Figure 5.

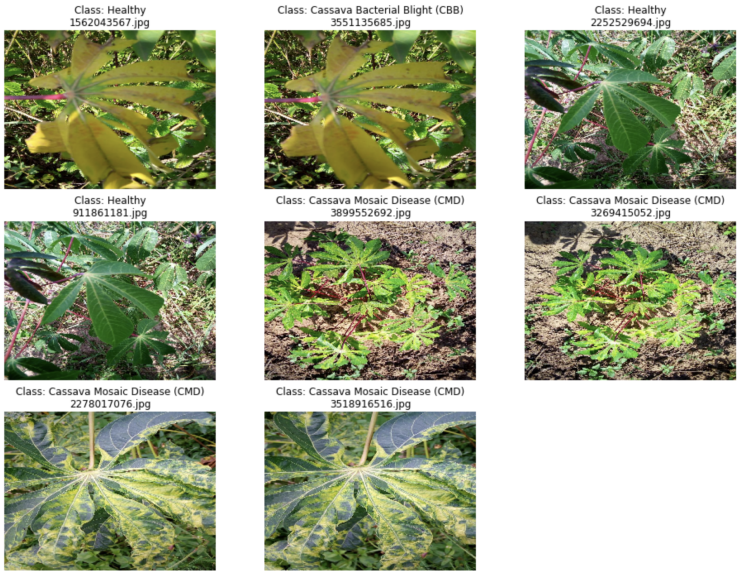

Figure 5: Duplicated images

SOLUTION: We removed the duplicates because it reduces the bias to the dataset and confusion to the model when faced to learn two or more same images with different labelling. It is possible to find even more duplicates by using more advanced similarity models for images but due to similarity of the over all images there was a lot of false positives for similarity models. I attempted to use CNN from imagededup python library but the similarity of the images drove to large amount of false positives when similarities were bellow a score of 90 percent. Above 90 percent I would get around 21 images as well like Hamming Distance model. But CNN is very slow, so due to time constraints and processing power we used Hamming Distance model instead.

5.0.5 High similarity within CBB/CBSD and CMD/CGM

Two diseases are so similar that false positive and false negative very high. There is 10.8 percent of false positives and 6.8 percent false negatives between CBB and CBSD. There is 7.6 percent of false positives and 4.9. We do not know if they are mislabels by the expert team but we noticed their similarity is creating confusion in the model. CBB and CBSD have a very similar leaf disease when CBSD disease covers the entire leaf. That makes it so similar to CBB. There are also cases of brown or gray spots on CBSD leaves which are not one of the side effects of CBSD disease from the research we did regarding the details of each disease symptoms on the leaf. SOLUTION: We did not changed the label from CBSS to CBB because that occurrence is very frequent and we did not find a specialist that could help us certify our findings regarding mislabeling of CBSS with CBB correct. We left the labels as they were done by the experts. However, if we solve this possible issue it should significantly improve accuracy. Our forecast is between 2 to 4 percent improvement on accuracy.

5.0.6 Images with far away focus

There are a lot of images that are fairway from the cassava plants' leaves or that is incorporating an entire plantation. Here is a demo video we did to show how the model only forecast healthy leaves when it is aimed far away from the leaves on a large plantation. `https://youtu.be/OWxVh_buiNg`

SOLUTION: The solution would be to remove images from far away and full plantation. Accuracy for such images are mostly predicted as Healthy as we showed in the DEMO VIDEO 2. We did not remove them in order to not severely affect our model against CropNet. Nonetheless, we believe such pictures are not helpful in training the model since the goal is to evaluate the disease of the leafs in a very close distance in order to avoid neighboring leaves. This is because it is fairly common for a leaf disease be surrounded by healthy ones or other diseases.

6 Modeling and result analysis

We used two model architectures, ResNext50 and CNN, while benchmark results were against MobileNetv3 used by CropNet. ResNext50 performed better than CNN and MobileNetv3. ResNext50 had an accuracy of 80.88 percent on a test set of 2123 images while CNN had an accuracy of 63.5 percent and MobileNetv3 had an accuracy of 77.6 percent. MobileNetv3 dropped from 88 percent accuracy with a skewed training set and validation test of 72 percent. With our fixed skewness and relabelling of our 2123 test set led to a drop of 10.4 percent in accuracy of CropNet. We successfully surpassed CropNet with our ResNext50 architecture by 3.22 percent. Those accuracies were also based on the following augmentations. CropNet's MobileNetv3 used the following augmentations: brightness, contrast, saturation, hue, and crop. Our ResNext50 used crop, transpose, rotation, shift rotation, pixel normalization, dropout, and cutout. CNN used brightness, contrast, saturation, hue, crop, rotation, and blurriness.

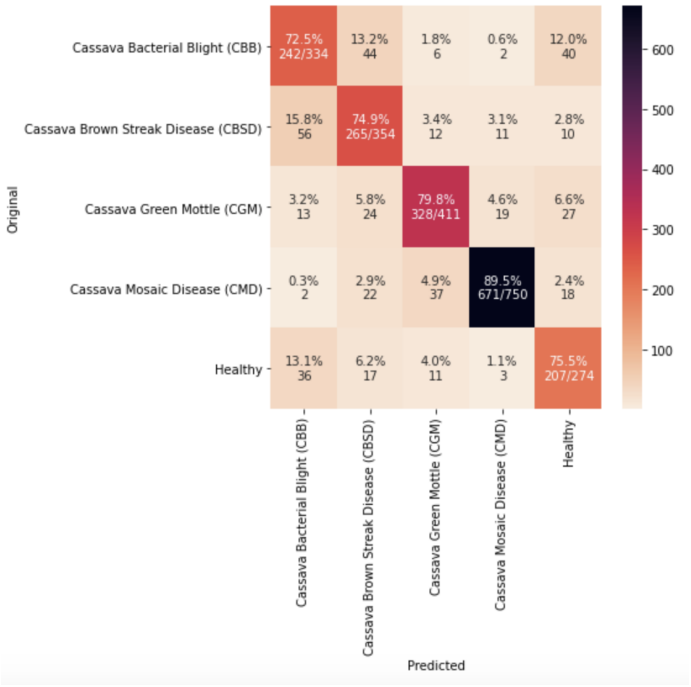

Figure 6: ResNext50 trained on 8460 images and tested against 2123 images

6.0.1 ResNext50

The result from our best model shows that relabelling, optimization and reduction of skewness best generalized the model and heavily reduced Healthy false negatives. We used 8460 images which was a much smaller fraction of Kaggle's 21397 data set seen on Figure 7.

6.0.2 ResNext50 trained on Kaggle's full data of 21397

The result our model architecture using Kaggle's full data set of 21397 without any relabelling from our side provided an accuracy bellow our model. This test gave a 78.2 percent which is slight higher than the CropNet, 77.6 percent, but lower than our model, 80.7 percent. This shows that our architecture is working better than CropNet's MobileNetv3 architecture. Moreover, our relabelling, data skewness reduction and optimization helped to drive better performance. However, this model showed some peculiar results. The Healthy data had a higher false positives but lower false negatives. CMD score was 96.5 percent as seen on Figure 7. That was a very accurate prediction for CMD. This shows that the more images we have for our model the more accurate it should become. However, CMD had around 13000+ images while all the other diseases had less than 3000 images. However, the price for high CMD and Healthy accuracy led to a very low CBB accuracy, 33 percent. That is way bellow 50 percent. Moreover CBB false positives was greater than true positives. This highlights even more the skewness of the data and miss-labeling from Healthy labeled data. We were able to fix the mislabeling on the Healthy data set. However, collecting 5x more images was a daunting task that would require extra monetary funding and time to get it done. That was not possible to accomplish in 3 months, so

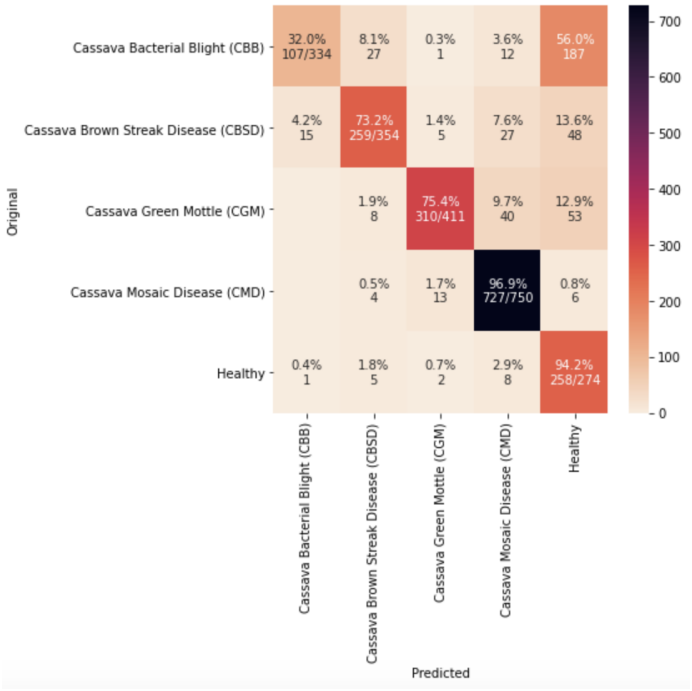

Figure 7: ResNext50 trained on 21397 images and tested on our 2123 images test set.

we focused on improving a smaller data set for a more even performance across all diseases. We achieved accuracy higher than 72 percent on all classes after our improvements.

7 Conclusion

We were able to successfully surpass CropNet by fixing mislabels, addressing skewness, parameterization and image augmentation. Despite reaching our goal of surpassing our benchmark, CropNet, we believe we did not reach results for MVP in order to launch this project as a product in the market. That conclusion came from long hours of research. As always what we first thought was not the most accurate one. We found out that we were only addressing one step of the Cassava's life cycle but that the camera OAK-D could address all the four life cycle steps. These life cycle steps are separated in planting, growth, cutting and storage cycle. Currently our leaf disease detection was only addressing the growth cycle of the Cassava. We found out that the biggest cause of transmission of the disease has been the mixing of stems during the cutting and storage process. Most of the disease's spread occurs with the contact of the stems when packed in pallets or cutting tools reused on healthy plants after usage on sick stems. This means that our model needs to learn on how to identify diseases on the Cassava stems and roots as well. Doing so would make the most usage of OAK-D camera while at the same time helping farmers to identify the diseases before the stem cutting process and storage. In other words, during the planting period we are unable to find out the stems that were already infected during the cutting and storage process because there are no leaves in them yet.

Figure 8: Steam nucleus health stage for planting

Moreover, we also found out that the circular shape of the stems' nucleus is paramount to determine the best period to initiate the planting cycle of Cassava [6]. As you can see on Figure 8 it is possible to build a model to recognize stem nucleus readiness for planting. On the other hand, root disease was the least important step in terms of saving the plant during its life cycle. Identifying the disease by uprooting the Cassava plant is unfeasible to save the plant because once uprooted it can't be planted back. However, commercially, root disease identification and shape of the root identification will help to determine the quality of the roots after its uprooting, washing and processing. To achieve our MVP goal, we would need to

improve our model by including stem's disease identification, stems nucleus identification and root diseases identification to increase the period usage of OAK-D camera. That would mean 100 percent utilization of the camera during the Cassava life cycle. Finally the last goal would be the implementation of monocular depth segmentation of the Cassava leaves by using the camera depth sensors to improve segmentation. However, in order to achieve the MVP goal of this project we would need the funding from the award of this tournament to be able to collect all the data from all life cycles of the plant in order to train it in our model. This way farmers would use OAK-D for the entire life cycle of the cassava plant which can drastically reduce disease transmission and increase even more production yield than we previously thought. We hope our solution is appealing and our goal to bring this to the market is worthy of an award to make it a reality.

References

[1] Tensorflow; CropNet Cassava Disease Detection; by Google, Makerere Artificial Intelligence (AI) Lab and National Crops Resources Research Institute (NaCRRI) In *URL: https://www.tensorflow.org/hub/tutorials/cropnet*$_cassava$

[2] Embrapa "First case of new cassava disease confirmed in Brazil"; by Research, Development and Innovation; published on 03/09/19; In *URL: https://www.embrapa.br/en/busca-de-noticias/-/noticia/46242292/confirmado-no-brasil-primeiro-caso-de-nova-doenca-da-mandioca*

[3] Sen Nag, Oishimaya; "Top Cassava Production Countries in the World", published on April 25 2017 in Economics; In *URL: https://www.worldatlas.com/articles/top-cassava-producing-countries-in-the-world.html*

[4] Matos, Canto, Patino, Souza, Schaun, Fukuda; "FARMER PARTICIPATORY RESEARCH: THE TURNING POINT FOR CASSAVA DEVELOPMENT IN NORTHEASTERN BRAZIL"; In *URL: http://www.fao.org/3/y5271e/y5271e07.htm*

[5] Maslov, Dimitry; OpenCV AI Kit OAK-1 Review and Custom Model Inference; Published on August 30, 2020. In *https://www.hackster.io/dmitrywat/opencv-ai-kit-oak-1-review-and-custom-model-inference-23ea3d*

[6] Suzane Nascimento; IBS Instituto BioSistêmico; Seleção de Manivas (Portuguese); Stem health photo on the minute 20:23/28:53 of the video. In *URL: https://www.youtube.com/watch?v=V9OMubDunbM*

[7] Makerere University AI Lab; "Cassava Leaf Disease Classification", Research Code Competition on Kaggle, In *URL: https://www.kaggle.com/c/cassava-leaf-disease-classification/overview*

[8] Mwebaze, Gebru, Frome, Nsumba, Tusubira, Omongo; "iCassava 2019 Fine-Grained Visual Categorization Challenge", published on December 24, 2019; In *URL: https://arxiv.org/pdf/1908.02900.pdf*

[9] Kaggle. Cassava Leaf Disease Classification. Makerere Artificial Intelligence (AI) Lab at Makerere University in Uganda., 2020. *URL: https://www.kaggle.com/c/cassava-leaf-disease-classification*