Final Report

Project 13: Counting stomata cell structures in plant leaves

Team Members:

Sihan Liu
Yan Wen
Xingyu Pan
Ruilun Wang
Zhengyu Wu

Introduction

The stomata are the key factor to control the growth of plants, it allows plants to absorb carbon dioxide needed for photosynthesis and help reduce water loss when it's hot or dry. Different trees have different properties of stomata, it is closely related to genes and environment and determined by both genes and environment. This also applies to the stomata on the surface of poplar leaves. For different gene types and different growth environments, the stomata of poplar trees may have different manifestations. This project investigated the relationship between the stomata number density, the stomata size, the genotype, the radiation and the planting location.

Background

The data we obtained is a stomata dataset consisting of 494 poplar trees that grow on the same experiment field, this project aims to analyze the difference of stomata properties related to different attributes. Stomatal properties include stomatal density, stomatal size and stomatal number. By analyzing the data and building the model to predict, we verified the correlation between stomata and these attributes.

Proposed Idea

2 In this project, the main topic has been separated into three research directions for a better and clearer analysis.

2.1 First Direction

Without considering the genotype, how the specific stomata properties correlated with each other. In this research direction, the main idea is to determine whether the stomata properties like size and number will be affected by the surface of the leaf, and also if there are any relationships between properties within the stomata. This direction can be splitted to three research questions:

2.1.1 Are there any differences of stomatal properties between two surfaces of leaf?

2.1.2 Are there any environmental factors affecting the stomatal properties?

2.1.3 Is there any correlation between stomata sizes and the counted number?

2.2 Second Direction

Different from the 2.1.1 above, in this direction, genotype will be enrolled in the research since the goal is to analyze the relationship between different genotype categories and stomata properties. If the genotype has relationship with stomata properties, then we push forward and do more specific analysis on how the genotype correlates with the stomata properties. Two research questions can be reached out from this direction:

2.2.1 Does the male parents of the plants affect the stomatal properties?

2.2.2 Does the irradiation level used to treat the seeds affect the stomatal properties?

2.2.3 How does the genotype determine stomatal properties on “up” and “low” surfaces?

2.2.4 Are there any correlations between image entropy and stomatal properties?

2.3 Third Direction

In the last direction, the target is to analyze whether variance of stomatal properties of replicated plants for a specific genotype will fluctuate greatly. Since there are lots of genotypes given in the dataset, this direction should enroll all the genotypes provided in the original dataset. This direction leads to one research question:

2.3.1 For a given genotype, how does the variance of stomatal properties among its different replicated plants correlate with the genotype itself?

Methodology

3.1. About the Data

3.1.1. Data Gathering and Interpreting

We sincerely appreciate Dr. Groover and his colleagues for the data they provided. The data we obtained are nearly one thousand images of leaf imprints taken under microscopes in which various numbers of stomata can be identified. The images come from nearly five hundred trees, half of them reveal the stomata on the “up” surface of one of their leaves, and the other half reveal the “low” surface.

For every image, we also received data about the number of stomata in each image. The stomata are annotated both manually by professional researchers and automatically by an online AI tool called “Stomata Counter”. The online tool also provides additional information about the quality of the image, such as entropy, etc.

Each poplar tree, along with its corresponding images, is labeled by a position (column, row) on the field, so we also have the data mapping all the positions of trees to their individual genotypes. For most of the trees, their genotype name includes three fields. Since these trees are the offspring of the same mother crossed with two distinct fathers, the first field (IFG or GWR) in their genotypes represents the father’s genotype, and “XXX” means it is unsure (interestingly,

the special genotypes, SO361SL and SO598SL, correspond to the two distinct fatheres). The second field is the irradiation level used to treat the seed to cause chromosomal breaks or not (0, 50, 100). The third field is the sequential number the researchers gave to each genotype within its larger classes. For every distinct genotype, there are on average three different replicated plants in our data.

3.1.2. Data Cleaning

A few images are of poor quality, such that there is no corresponding data for it. Therefore, the data of one surface could be missing for a few trees. To ensure the completeness of data for every single tree and for the convenience of future analysis, we only keep the trees with a complete set of data for further use.

3.1.3. Obtaining the Size of Stomata

Although the number of stomata in each image is manually counted, and the Stomata Counter tool also provided image quality information, we still don't have any data about the size of stomata. Due to the limitation of image qualities and the sophisticated techniques required, calculating the size of every single stoma in all images is not achievable. Hence, we developed an easy method to estimate their average size in an image using the OpenCV library.

First, we convert the original image into a grayscale one and preprocess it to improve contrast. After this step, we use adaptive thresholding to convert it into a binary image such that the stomata and other tissue structures are in white pixels. At the moment, a stoma is still detected as a few curves which sketch its contour. It would be convenient to first convert the stomata to blobs, so that the area (or size) can be easily calculated. To achieve this, we perform dilation and erosion on the binary image. After this step, we find all the contours in the image and calculate the area enclosed by each of them. Areas that are too large tend to be other noticeable tissues detected by the previous operations, whereas areas that are too small tend to be noise and unnoticeable tissues in the background. Holding the parameters in the functions unchanged, we ran the script on several sample images. We found that the size of a typical blob of stoma, calculated by its contour, has an upper bound of about 3500 white pixels.

From what has been discussed above, our resolution is to calculate the average value of the largest 20 contours that are smaller than 3500 pixels and use this number to represent the average value of stomata in the image. Of course, this can be considered as an overestimation. But there are two reasons that suggest that we should not use the smaller values. First, the probability of counting in noise or other tissues would be higher if we were to do so. Second, even if the blob being counted is truly a stoma, there is a very large chance that it's poorly detected so its blob is not plump enough. Thus, although being overestimating, the data should have high representativeness.

3.1.4. Constructing the Dataframes

Using the data we obtained from the previous sections, we are now able to compute the density of stomata. Since the assumption is the images are of the same size and the same magnification when they were being processed, we can simply use the product of number and size to represent an estimation of the stomatal density and a division over image size is not needed. We finally constructed two main dataframes to work with for our convenience.

The first data frame has each of its rows correspond to a different tree. The columns are information about the counted number of stomata and their average size, obtained from the two images of both surfaces of its leaf.

The second data frame has each of its rows correspond to a different genotype. The columns are the data that describes the difference across individual replicates of the same genotype.

3.2. Research on the Data in General

First, we want to find out how the stomatal properties (namely, manually counted number, estimated size and estimated density on two surfaces) are distributed in general, or what's the relationship between them and other factors other than genotypes.

3.2.1. Research Questions (1)

First, we would like to analyze the difference of stomatal properties between the two surfaces.

To visualize the data and see the distribution of them, we plot box plots for the three quantities on both leaf surfaces and compare them. Then we can use these quantities to compute the differences we are interested in for every plant (for every row in the first data frame).

3.2.2. Research Questions (2)

We try to find if there are environmental factors affecting the properties of stomata. The only information we have in terms of environment is the plant's position on the field.

To answer the question, we plot the stomata properties of all plants distinguished by rows and columns respectively and see if there is some pattern of distribution along the rows or columns. To prove the possible correlations, we construct separate regression models and fit them on the average values of stomatal properties versus the column / row numbers

3.2.3. Research Questions (3)

After computing the sizes of stomata, we are interested in if there is a correlation between this property and the counted number of stomata.

We plot scatter plots with counted numbers on the horizontal axis and sizes on the vertical axis. After inspecting the plots we decided that there was no need for a regression model in answering this question.

3.3. Research on Differences across Genotypes

Now we are interested in how different genotypes affect the stomatal properties. Since they are just acquired as pure numbers, the analysis techniques used for these three properties are identical. The main data frame to work with in this section is the first one mentioned in 3.1.4.

3.3.1. Research Questions (4)

This subsection discusses the method in finding how stomatal properties may be affected because the plants differ in their male parents (which is indicated by the first field in the name of a specific genotype).

We first draw box plots of stomatal properties of all plants, classified by the first field of their genotype names, and we observe the distributions of data. The mean values are also calculated at the same time. Then we build three decision tree models, one for each property, to predict the first field in the plant's genotype (GWR of IFG). Notice that we discarded all the plants with an unknown male parent (XXX). Take the property "manually counted number" as an example, the input features for the model would be the number of stomata on the "up" and "low" surfaces, the average and the difference of them.

3.3.2. Research Questions (5)

This subsection is similar to the previous one, except now we are interested in how the irradiation level used to treat the seeds may affect the plants stomatal properties.

Similarly, we plot the data and calculate the means first, and then we fit decision tree models. Notice that this time the XXX's can be included since we have the irradiation level for them. However, we identified only 15 plants whose seeds were treated with 0 irradiation level and 7 plants for 0 irradiation level. To produce a meaningful accuracy of our model, the test set being used are constructed with all these 22 plants, with another 15 plants sampled from the 100 irradiation level class.

3.3.3. Research Questions (6)

This subsection discusses the method in finding how stomata counted number, estimated stomata size and estimated stomata density on the "up" and "low" surfaces may be determined by genotype using the model of neural network.

We first create two data frames where each of their rows only correspond to a different genotype of plants with the known male parent(IFG, GWR), the columns describe stomatal properties(cross, radiation, sequence number, manual counted number, size, density, etc,.) of the tree with the corresponding genotype. Since each genotype could have at least two trees and we only need one tree's stomatal properties, then we get all trees with the same genotype and the stomatal properties of two of the trees are selected at random to put in the first data frame and second data frame separately. Therefore, the two data frames contain a total 183 genotypes and corresponding 183 stomatal properties of the trees. After finishing the creation of two data frames, we start to build the neural network models with Keras, to

predict the number of stomata, the size of stomata and the density of stomata on the up and low surfaces, meaning that we will have 6 neural network models in total. We use the first data frame as the training set and second data frame as the testing set. For each model, the input features are the male parent (IFG or GWR), irradiation level(0, 50, 100) and sequential number and one output variable. To get better training performance, we apply one-hot encoding to encode the features. We create sequential models in Keras and use a fully-connected network structure with four layers. Then we compile the model with the MSE loss function and the optimizer of stochastic gradient descent algorithm. Finally we observe the R-squared values returned by all 6 neural network models to decide whether there is a strong relationship between genotype and stomata properties.

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_4 (Dense)              (None, 64)                8128
_________________________________________________________________
dense_5 (Dense)              (None, 128)               8320
_________________________________________________________________
dense_6 (Dense)              (None, 128)               16512
_________________________________________________________________
dense_7 (Dense)              (None, 1)                 129
=================================================================
Total params: 33,089
Trainable params: 33,089
Non-trainable params: 0
```

3.3.4. Research Questions (7)

During the process of estimating the average size of stomata, we observed that higher contrast of the image results in more plump shapes of detected stomata, so the detected size is larger. Thus, it is useful to find any possible correlation between the genotypes and the image contrast, which can be measured by its Shannon entropy.

The related data can be found in the original files provided by Dr. Groover, and we investigate the possible relationships by plotting all the image entropy values, classified by the plant genotypes of the corresponding images.

3.4. Research on Differences within Genotypes

In this section, we want to investigate the variance of stomatal properties among the replicated plants of the same genotype. The main data frame to work with in this section is the second one mentioned in 3.1.4.

3.4.1. Research Questions (8)

This subsection is mainly about studying how some properties of stomata intertwined with each other in the same genotype. The stomatal properties are considered as the possible factors. We made two

plots, and the first plot is between density differences and genotype. The second one is between manual count difference and the genotype.

After finishing these plots, it is clear that there is no need to build regression models to answer these questions.

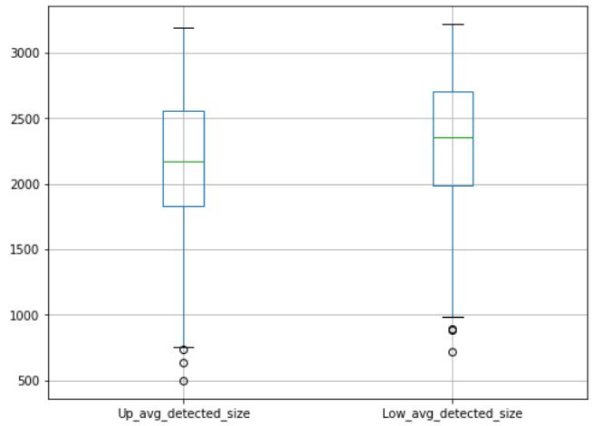
Results & Discussions

4.1 Result & Discussions of Research Questions

This section discusses the results of our methods, with respect to the eight research questions.

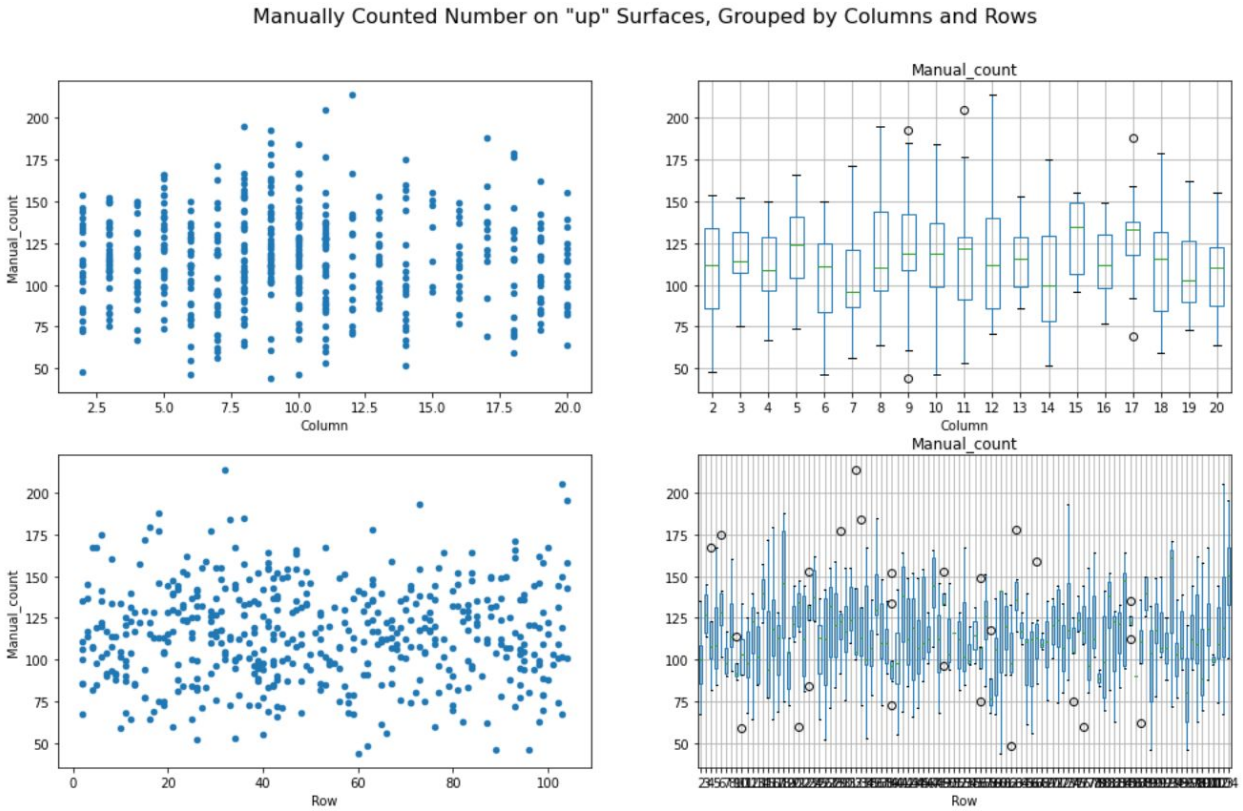
4.1.1 Result of Research Question 1:

We observed from the box plots that “low” surfaces have higher median, Q1 and Q3 values than “up” surfaces, and this is true for all three stomatal properties. Displayed below is the box plot for estimated size of stomata.



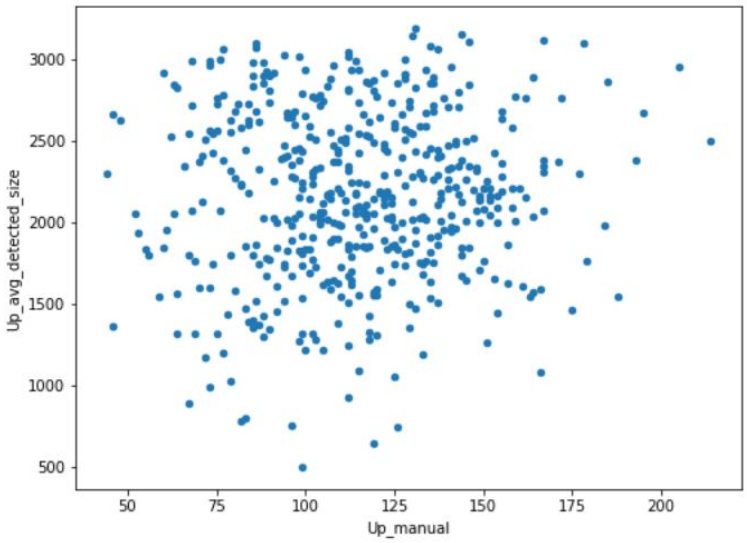
4.1.2 Result of Research Question 2:

The plots show that there is no observable correlation between all stomata properties and the location of the plant on the field. The R-squared values of the regression models are all zeros, and the various p-values for different predictors are all greater than 0.05. Here is one example of our results on the plot, other results are similar.



4.1.3 Result of Research Question 3:

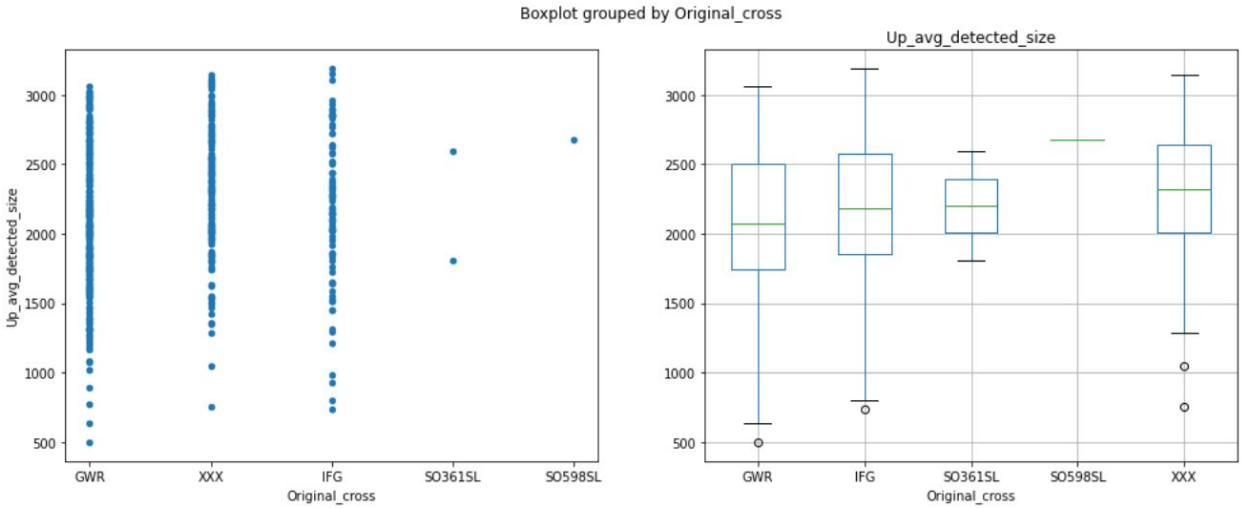
Below is the resulting scatter plot (Up_manual = manually counted number of stomata in an image, Up_avg_detected_size = the estimated average size of stomata in an image) in which no type of correlation or trend can be observed. With the randomness clearly presented, there is no need for a regression model to analyze any correlation.



4.1.4 Result of Research Question 4:

For all stomatal properties, the data distribution of IFG and GWR trees almost completely overlap with each other such that it's almost impossible to distinguish these two classes given any stomata properties. The plot of stomata sizes, grouped by the first field of genotype names, is shown below. The plot also includes the unknown XXX's and the two fathers for comparison.

Trained on all rows, the three decision tree models all have an average accuracy less than 0.7 (randomness is involved during splitting train and test sets) when tested on the test set. This result is not meaningful, because after discarding the XXX trees, 75% of the rest are in the GWR class. We can conclude that the decision tree models do not perform better than a simple model that just always predicts the GWR class. Hence, there is no evidence for any correlation between the father genotype of plants and their stomatal properties.



4.1.5 Result of Research Question 5:

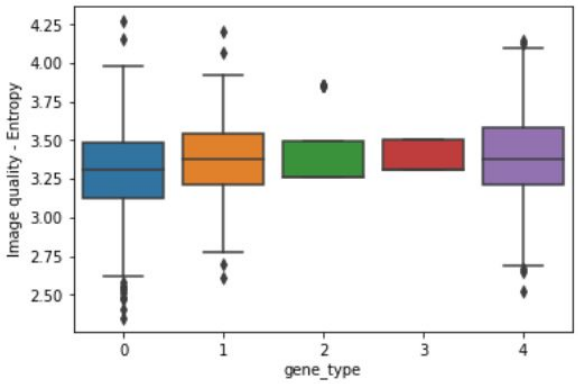
Similarly with Research Question 4, by plotting the graph, it is clear that there is no obvious relationship between stomata properties and the irradiation level. However, it is worth mentioning that there is a limitation when generating the irradiation level related models since the amount of irradiation data is far less than the stomata data. Also, the distribution of irradiation data is uneven.

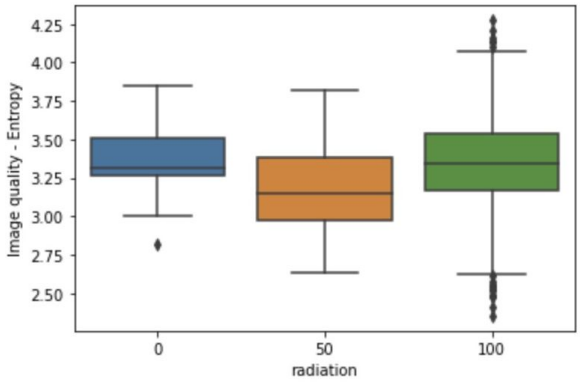
4.1.6 Result of Research Question 6:

We generated a neural network model to analyze how the stomata properties on the “up” and “low” surfaces may be determined by genotype. In this model, the R-squared score of up surface stomata counted number (manual_up) is about 0.113, the score of low surface stomata counted number (manual_low) is about 0.314, the up surface stomata size (up_size) is about -0.087, the low surface stomata size (low_size) is about 0.0719, the up surface stomata number density (up_density) is about 0.103, and the low surface stomata number density (low_density) is about 0.025. Since all the R-square scores are lower than 0.5, we can conclude that there is no evidence to show the stomata properties on whatever the “up” or “low” surface are correlated with the genotype categories. Which means there is no relationship between genotype categories and the stomata properties on leaf surfaces.

4.1.7 Result of Research Question 7:

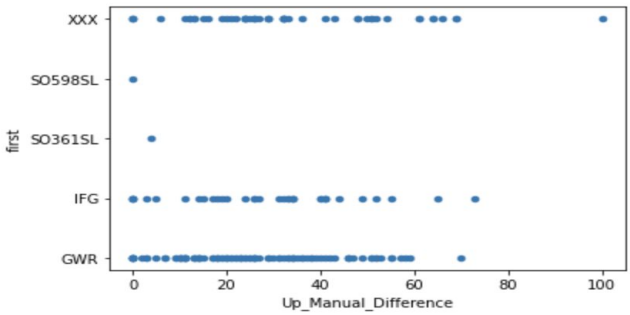
It can be seen from the two box plots that the distribution of image entropy, grouped by the first and the second fields of genotypes respectively, overlap significantly with each other. We therefore conclude that there is no evidence of correlation between the first two fields of genotypes and the image entropy or contrast.





4.1.8 Result of Research Question 8:

Since on average there are three trees in each genotype category, it is important to figure out the average stomata properties difference among those three trees in each category. By generated models with related data, such a result comes out that the stomata property variance within a given genotype are not correlated with the genotype categories. This means there is no evidence of relationship between stomata properties difference and genotype categories.



4.2 Some additional findings

During the process of answering the research questions, we found two noticeable facts. The first one is that although IFG and GWR classes have greatly overlapped distributions of their stomatal properties, the IFG class has slightly greater median, Q1 and Q3 across all stomatal properties than GWR. The second is that the XXX class, which should be an unknown mix of IFG and GWR plants, has even higher values in the stomatal properties.

We hope that the problem can be addressed by obtaining more data from Dr. Groover in the future.

Limitations

Based on the dataset provided by our mentor Dr. Groover, our group generated some models to analyze the relationship between variables within the dataset and the number density of stomata. However, beyond those variables, the density of stomata on a leaf might be affected by other factors like the global warm effect and some chemical elements that current technology can not detect. That is, the analysis we produced in this project can only reflect the relationship between factors the botanists have already found for the stomata. We can not conclude whether the stomata density will be affected by the other factors that are not included in the dataset we used since this behavior is inaccurate and unreliable.

Since the radiation data amount is far less than the amount of stomata data we used to build up a model for generating relationships and the distribution of radiation data itself is also obviously uneven, the accuracy of the radiation model may have errors reflecting the relationship between stomata density and the radiation degre. Thus, it is necessary to include more data to support the influence of this variable on the accuracy of the model.

Furthermore, during our research, such a fact has been found that the quality of those stomata images provided by Dr.Groover might affect the accuracy of data. When generating the stomata amount on images, low quality images can cause the automatic counting tool we used to confuse the stomata and noise, and affect the accuracy of models in future analysis. Also, the contrast also affects the accuracy of models we generated with the amount of stomata on the image. Images with high contrast will generate more stomata than the low contrast images, and it will also limit the accuracy of models. Therefore, all the models might be affected by this limitation, and we are unable to solve this limitation since all the images are not taken by our groups.

Conclusion & Future Work

This paper has presented an analysis of the relationships between stomata properties and its genotype. The paper presented a varied range of models that analyze and explain how stomata properties are affected by factors enrolled in its growing environment. Provided mathematically evidence to show the stomata properties are not related to its genotype categories and irradiation level. The methods presented herein are highly reproducible, and the results are reliable thereby. We can improve the performance of decision tree models if there is more genotype data with irradiation levels in 0 and 50. To improve the accuracy of the Online stomata counter AI, we can train it further by uploading the stomata images which we have processed with manual count.

Works Cited

Hetherington, Alistair M., "The role of stomata in sensing and driving environmental change", Department of Biological Sciences, The Lancaster Environment Centre, University of Lancaster, Lancaster LA1 4YQ, UK, 2003. https://www.researchgate.net/profile/Alistair_Hetherington/publication/10603718_Hetherington_A_M_Woodward_F_I_The_role_of_stomata_in_sensing_and_driving_environmental_change_Nature_424_901-908/links/00b495241cd5ae3543000000/Hetherington-A-M-Woodward-F-I-The-role-of-stomata-in-sensing-and-driving-environmental-change-Nature-424-901-908.pdf

Koho, S., Fazeli, E., Eriksson, J. *et al.* Image Quality Ranking Method for Microscopy. *Sci Rep* 6, 28962 (2016). https://doi.org/10.1038/srep28962

TETOUHE KILIMOU, Edouard. "Images Processing: Segmentation and Objects Counting with Python and OpenCV". 2019. https://medium.com/analytics-vidhya/images-processing-segmentation-and-objects-counting-in-an-image-with-python-and-opencv-216cd38aca8e