HIGH SCHOOL EDITION
Journal of Student Research
Volume 12 Issue 1 (2023)

Plant Toxicity Classification by Image

Eera Bhatt[1] and Clayton Greenberg[#]

[1]Troy High School
[#]Advisor

ABSTRACT

Poisonous and non-poisonous plants have extremely similar visible features to any non-botanist, which puts those in danger who are frequently present in areas containing various plants. Failure to distinguish harmful plants from safe ones puts several people at a high risk of accidents and potential health issues after contact with a toxic plant. While previous work has found ways to classify specific types of plants, a limited amount of research has been done on toxic and non-toxic plants of several species. Since differentiating between dangerous and safe plants is a complex task for a human brain, this study approaches the issue through machine learning models starting with a convolutional neural network (CNN) and discovering that a logistic regression model—trained on a dataset with manually designed features—has the best performance with the dataset used. The neural network demonstrated overfitting which was likely caused by the inconsistent backgrounds of images within the dataset. The logistic regression model achieved an accuracy of 97.37% in its predicted classifications for the plants. The best-performing logistic regression model contained the three-leaf and dark red stem features indicating that these two features are the most reliable ones used to distinguish between toxic and non-toxic plants.

Introduction

Several poisonous plants are present around homes and work areas that may irritate one's skin and lungs if absorbed, inhaled, or physically touched. In some cases, poisonous plants are often unavoidable outside and there are often no reliable ways to distinguish between toxic and non-toxic ones. Moreover, the lack of ability to differentiate between safe and harmful plants puts several people at a risk of accidents and serious health complications. As plant identification involves complex patterns, a supervised machine learning model can be created to effectively distinguish between toxic and non-toxic plants and identify the most prominent features in each class of plants. The data used throughout this study included vision data for deep learning and numerical data for a logistic regression model. In our deep convolutional neural network (CNN), an image sample for each plant was inputted into the network and the output was the plant's predicted classification as toxic or non-toxic. In the logistic regression model, multiple features of each plant sample were inputted, and the output was the plant's predicted toxicity label.

Background

While toxic plant classification is not the most common issue addressed using artificial intelligence, a few machine learning approaches have been tested in the past, particularly the convolutional neural network (CNN) for the image-based classification of plants. However, the samples that have been classified in previous work have consisted of extremely specific species of plants. Some research has dealt with classifying mushrooms in particular as edible or poisonous (Ketwongsa et. al, 2022) while other work has involved the differentiation between certain toxic and non-toxic herbs (Cho et al., 2019). These deep learning methods have had high accuracy percentages, especially with clean datasets of all plants facing the same direction in the same background which has helped the classification models accurately identify patterns in the vision data. However, the classification of common toxic plants of various species

 Volume 12 Issue 1 (2023)

has been an area not explored in depth thus far. The aim of this work was to find a model that could accurately classify common plants by their toxicity with a high sensitivity, or a high number of positive predictions compared to the total number of positive samples.

Dataset

The deep learning model used for classification uses vision data and contains images whereas the logistic regression model was based on numerical data used for a binary classification. The image data used in the models are scraped from a search on the iNaturalist site for each plant's scientific name. iNaturalist is a public networking site through which scientists share their observations about nature. A clean subset of the original dataset was used which contained 95 images of toxic plants and 95 images of non-toxic plants, each with its respective "toxic" or "non-toxic" label in addition to a species name. As the dataset contained images of multiple sizes, pre-processing for the convolutional neural network (CNN) included cropping each image to its 300 pixels at the center, as shown in the image below.

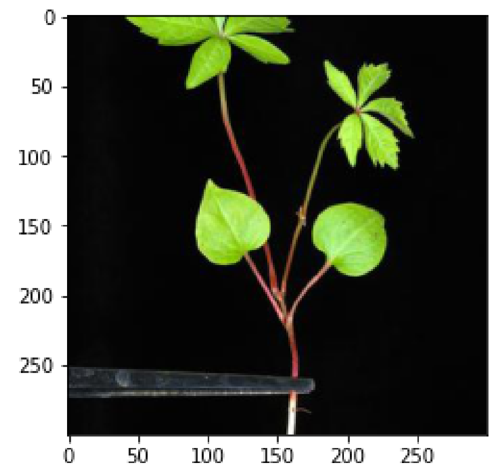

Figure 1. A pre-processed plant image sample is cropped such that only the 300 pixels at the center remain.

Within the cleaner subset of the data created for the models, five species of non-toxic and three species of toxic plants were present for a total of eight species. While there was an imbalance in the various species of the plants in this study, an equal number of toxic and non-toxic samples were still included in the final subset.

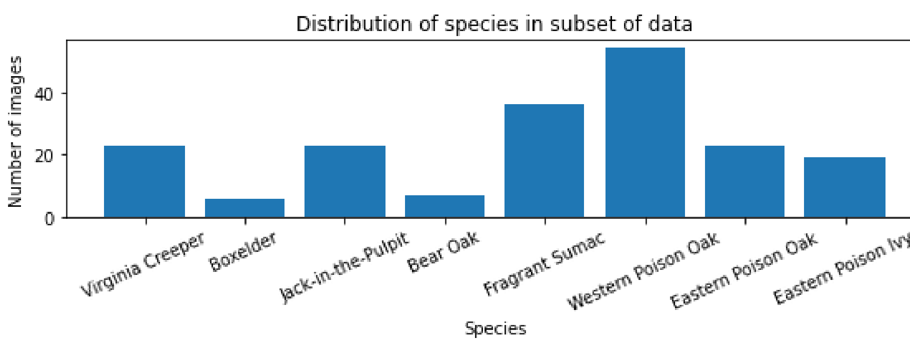

Figure 2. A bar graph showing the distribution of samples among the eight species used in the clean data subset.

Methods

The logistic classification model used original data produced by the authors of this paper based on certain features of each iNaturalist image. The features chosen for the models were selected based on prior reading material comparing toxic and non-toxic plants as well as from visual observations after browsing the samples. The features we incorporated into the data included: *shiny, 3-leaf, white-flower, green-flower, thorns*, and *dark red stem*. We determined these features using visual differences in the toxic and non-toxic samples that were visible to the human eye.

Figure 3. A raw sample (prior to being cropped) picture from the dataset with a true binary value for the *shiny* feature.

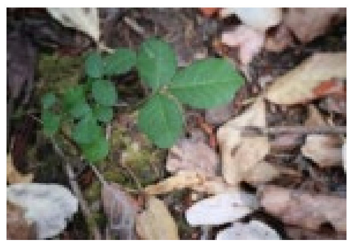

Figure 4. A raw sample (prior to being cropped) picture from the dataset with a true binary value for the *3-leaf cluster* feature.

Figure 5. A raw (prior to being cropped) sample from the dataset with a true binary value for the *white-flower* feature.

Figure 6. A raw (prior to being cropped) sample from the dataset with a true binary value for the *green-flower* feature.

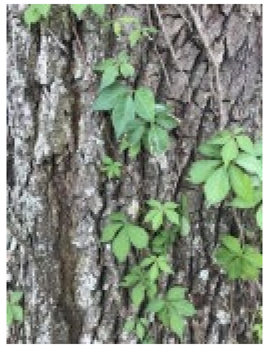

Figure 7. A raw (prior to being cropped) sample from the dataset with a true binary value for the *thorns* feature.

Figure 8. A raw (prior to being cropped) sample from the dataset with a true binary value for the dark red stem feature.

The Convolutional Neural Network

The issue of toxic plant classification was addressed by running a convolutional neural network (CNN) to classify images and exploring feature combinations for the images in a logistic regression model. The convolutional neural network (CNN) is a deep learning algorithm used for image data in which multiple layers process an image and recognize complex shapes within specific regions before outputting a classification. For the neural network model, the data was inputted into a *sklearn* model and split with a training size of 0.8 and a test size of 0.2 before training the CNN. We tried training for three, five, and ten epochs.

Logistic Regression

Logistic regression is a common machine learning algorithm used for the classification of data into multiple categories. Logistic regression is commonly used for models of binary classification that involve sorting data into two categories based on features. The logistic regression model implemented to classify toxic and non-toxic plants used features that were based on the image data in the CNN. The initial features in the dataset included being 3-leaf, shiny, thorns, dark red stem, and growing near white flowers or green flowers. A *sklearn* model was used for the data which was split into portions of 80% train and 20% test data.

Results and Discussion

The performance of the convolutional neural network (CNN) showed no change based on the number of epochs that were run. While the accuracy of the model achieved above 80% for training data, the performance scarcely exceeded a random chance accuracy on the validation data.

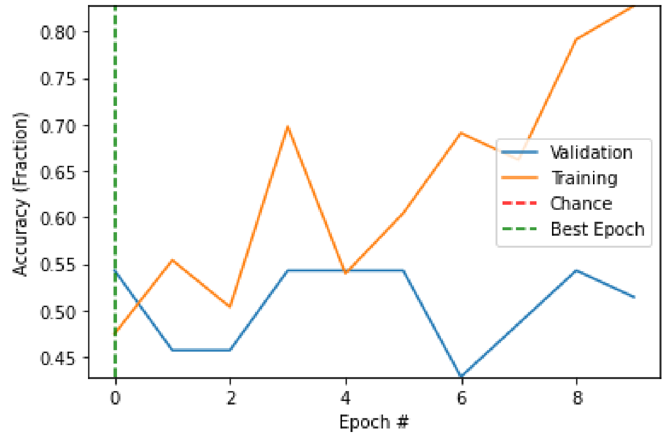

Figure 9. A line graph showing the accuracy of the convolutional neural network (CNN) on the training and validation data.

The prominent decrease from the training data to the validation data accuracies demonstrated large amounts of overfitting which was undesirable as the model was likely to make inaccurate predictions on future plant data. In rare cases during which the test data accuracy exceeded 50%, the convolutional neural network (CNN) displayed an unwanted bias towards the class that happened to be overrepresented in the current 20% testing sample.

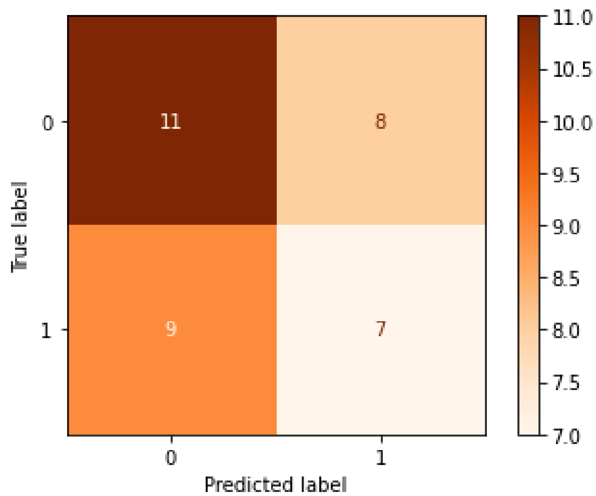

Figure 10. A confusion matrix for the convolutional neural network (CNN) performance on the test data.

Figure 10 indicates that the greatest number of outputs contained accurate predictions of non-toxic plants. According to the confusion matrix, approximately 57.14% of the data was predicted as being non-toxic while the remaining 42.86% of images were classified as being toxic. While in a few runs of the neural network, these non-toxic outputs slightly increased the accuracy, a bias toward toxic predictions is ideal for the best-performing model as the purpose of toxic plant classification is to identify the most harmful plants and avoid close contact with them.

After the poor accuracies and undesired biases, images within the subset of data were replaced to include more variety of color and features in the toxic plant category. This filtering process was implemented in hopes that the model would show a favorable bias toward toxic plant predictions. However, negligible changes were visible in the results even after the creation of a clean dataset.

The logistic regression model picked up on certain patterns which weren't caught through the convolutional neural network (CNN) and the result was a dramatic increase in the model's accuracy. Among the six features used in the original data, *white-flower, green-flower,* and *thorns* were removed as they hindered the accuracy of the model. The two logistic regression feature combinations with the best performance both achieved accuracies of 97.37% and the final decision for the most preferred model depended on feature selection. One of the two classification models included the features of *3-leaf, shiny,* and *dark red stem* while the alternate model used only *3-leaf* and *dark red stem* as features.

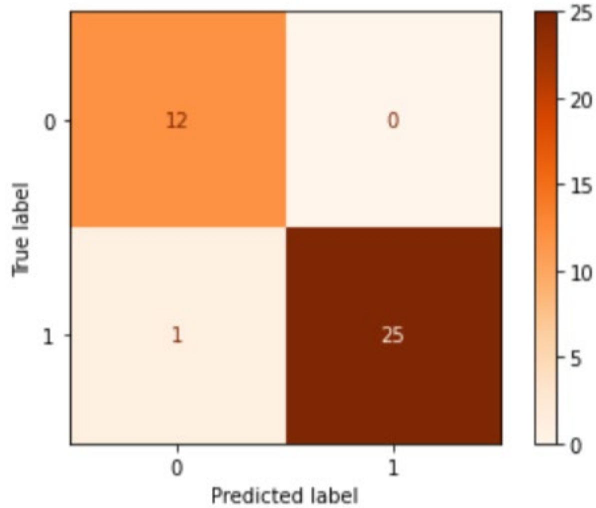

Figure 11. Performance of the logistic regression model using *the dark stem, 3-leaf cluster,* and *shiny* features to predict each sample's toxicity.

As displayed by *Figure 11*, the model that used three features had an accuracy of 97.37% and sensitivity percentage of 96.15% meaning that most truly toxic plants were identified as toxic, but not all. Sensitivity refers to the ratio of true positive predictions to all actual positives in a confusion matrix.

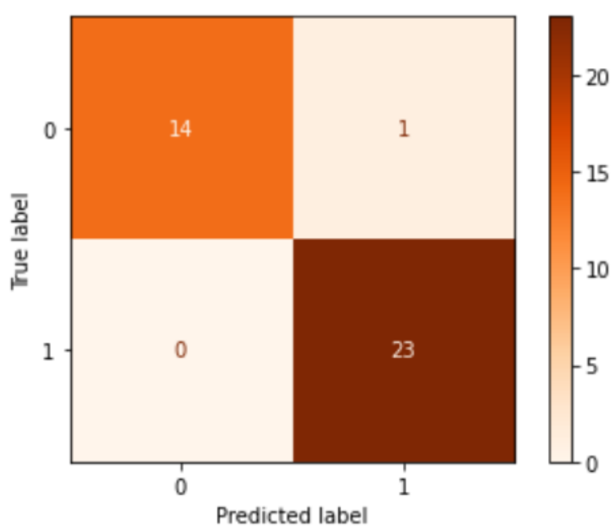

Figure 12. Performance of the logistic regression model using only the *dark red stem* and *3-leaf cluster* features to predict each sample's toxicity.

Meanwhile, the model referred to in *Figure 12* achieved the same high accuracy of 97.37% with a high sensitivity of 100%. The fact that all toxic plants were accurately classified as toxic indicates that the two-feature model's slight bias is toward identifying positives. The higher accuracy and sensitivity percentages of the two-feature logistic regression model makes it outperform all previous models discussed in this work. The results also suggest that the color of each plant's stem and its number of leaves in a cluster are the largest indicators of the plant's toxicity.

Conclusion

The final logistic regression model created using the *3-leaf* and *dark red stem* features effectively classified the harmful plants as toxic, which is the primary aspect of a successful model in this study. However, the models used in this study have a few limitations. For instance, the number of 190 samples makes the dataset relatively small as compared to those used in most machine learning classification algorithms. Additionally, the deep convolutional neural network (CNN) might have had a decreased performance due to the inconsistent nature of the iNaturalist dataset; the backgrounds for each plant image sample were different from one another, which likely interfered with the neural network's ability to detect complex relationships in the data. The most considerable drawback is that for the model to classify the plants effectively, a human was needed to annotate the data with the number of leaves in each sample. However, the model analysis shows that dark red stems and 3-leaf clusters are the most successful features to use in future work when detecting toxicity in plants. With a potentially larger and cleaner dataset, a fresh deep learning model will be promising for related future studies as neural networks are able to identify particularly abstract patterns in images.

References

Cho, J., Jeon, S., Song, S., Kim, S., Kim, D., Jeong, J., Choi, G., & Lee, S. (2019). Identification of toxic herbs using deep learning with focus on the sinomenium acutum, Aristolochiae manshuriensis caulis, akebiae caulis. *Applied Sciences*, *9*(24), 5456. https://doi.org/10.3390/app9245456

Ketwongsa, W., Boonlue, S., & Kokaew, U. (2022). A new deep learning model for the classification of poisonous and edible mushrooms based on improved alexnet convolutional neural network. *Applied Sciences*, *12*(7), 3409. https://doi.org/10.3390/app12073409

Ng, A., Tandon, S., Wang, T., Huval, B., Hannun, A., Maas, A., Coates, A., Suen, C., Mai, Y., Foo, C. Y., & Ngiam, J. (n.d.). Unsupervised Feature learning and Deep Learning Tutorial. Retrieved November 22, 2022, from http://ufldl.stanford.edu/tutorial/

Yalcin, H., & Razavi, S. (2016). Plant classification using Convolutional Neural Networks. *2016 Fifth International Conference on Agro-Geoinformatics (Agro-Geoinformatics)*. https://doi.org/10.1109/agro-geoinformatics.2016.7577698