Using Deep Learning for Detecting Soybean Diseases

Prithviraj Lakkakula

3/9/2022

Contents

Unstructured Data

Unlike traditional structured data, images analysis and the text data are considered as unstructured data. The analysis of unstructured data involves first converting unstructured data into structured data and then proceed with the analysis.

Deep Learning and Keras

Deep learning is a subfield of machine learning that uses neural networks. A simple neural network consists of an input layer, a hidden layer, and an output layer. The term 'deep learning' is used to model neural networks that has more than one hidden layer.

Keras is a high-level neural network application programming interface (API) for deep learning . It uses Tensorflow by Google as a backend. A front end is a user interface (what the user sees) and a backend is a server application and database that works behind the scenes to deliver the information to the user.

Soybean Disease Image Data

In this post, I will illustrate image classification and recognition using Keras package with 20 images for each of the four diseases of soybean. The four soybean diseases include bacterial blight (BB), bacterial pustule (BP), downy mildew (DM), and sudden death (SD) syndrome. In other words, this analysis is essentially a deep learning supervised approach that involve labeling of soybean disease images. A total of 80 disease images will be used in this illustration.

Before proceeding, first we need to download and call the libraries of the following packages. Please follow the steps below.

Step 1. Loading required R packages

```
#The line of R code that starts with '#' is a comment. For example, this is a comment.
#install.packages("BiocManager")
#BiocManager::install("EBImage")
library(EBImage) #EBImage, an R package used to handle and explore image data
library(keras)   #Keras is a high-level neural network API for deep learning
```

```
##
## Attaching package: 'keras'
```

```
## The following object is masked from 'package:EBImage':
##
##     normalize
```

Step 2. Read Images

The following chunk of R code reads all soybean disease images. In our case, it is a total of 80 disease images for our illustration purposes.

```
#setwd('/Volumes/RAJ/DLImages/idata')
images = list.files(pattern="*.JPG")
myimages <- list()
for (i in 1:length(images)) {myimages[[i]] <- readImage(images[i])}
```

Step 3. Exploring Soybean Disease Images

In the following chunk of R code, the *print* function provides an output that converts unstructured data, that is image, to structured data (numbers). In other words, the dimensions of the image is converted into data points (**pixels**). The **print** function provides an output that consists of dimensions (dim) for each of the four diseases. First observation is that the first and last images are of different size and second and third image are of the same size. The **dim** in the output consists of three numbers. For example, if you take the first image, the dimensions are **6016 times 4016 times 3** which when multiplied gives **72,480,768** pixels as shown in the histogram figure of first image. The number 6016 is width of that image, 4016 is the height of the image, and 3 indicates the number of channels. In our case, as we are dealing with images in color, the number of channels are 3, indicating RBG (red, blue, green). If it were a grayscale image, the last value in the **dim** would take a value of 1 (not 3).

```r
m <- c(1, 21, 41, 61)
for (i in m) {print(myimages[[i]])}
```

```
## Image
##   colorMode    : Color
##   storage.mode : double
##   dim          : 6016 4016 3
##   frames.total : 3
##   frames.render: 1
##
## imageData(object)[1:5,1:6,1]
##           [,1]      [,2]      [,3]      [,4]      [,5]      [,6]
## [1,] 0.4352941 0.4313725 0.4235294 0.4117647 0.4000000 0.3960784
## [2,] 0.4392157 0.4352941 0.4274510 0.4156863 0.4039216 0.3960784
## [3,] 0.4392157 0.4352941 0.4274510 0.4196078 0.4078431 0.4000000
## [4,] 0.4235294 0.4235294 0.4235294 0.4196078 0.4117647 0.4078431
## [5,] 0.4196078 0.4156863 0.4196078 0.4235294 0.4156863 0.4078431
## Image
##   colorMode    : Color
##   storage.mode : double
##   dim          : 4288 2848 3
##   frames.total : 3
##   frames.render: 1
##
## imageData(object)[1:5,1:6,1]
##           [,1]      [,2]      [,3]      [,4]      [,5]      [,6]
## [1,] 0.7921569 0.7921569 0.7960784 0.7960784 0.8000000 0.8000000
## [2,] 0.7882353 0.7921569 0.7921569 0.7921569 0.7960784 0.8000000
## [3,] 0.7882353 0.7921569 0.7921569 0.7921569 0.7960784 0.8000000
## [4,] 0.7882353 0.7921569 0.7921569 0.7921569 0.7960784 0.8000000
## [5,] 0.7882353 0.7921569 0.7921569 0.7921569 0.7960784 0.7960784
## Image
##   colorMode    : Color
##   storage.mode : double
##   dim          : 4288 2848 3
##   frames.total : 3
##   frames.render: 1
##
## imageData(object)[1:5,1:6,1]
##           [,1]      [,2]      [,3]      [,4]      [,5]      [,6]
## [1,] 0.5647059 0.5607843 0.5607843 0.5607843 0.5647059 0.5647059
## [2,] 0.5725490 0.5686275 0.5647059 0.5568627 0.5647059 0.5686275
## [3,] 0.5803922 0.5764706 0.5725490 0.5647059 0.5686275 0.5686275
## [4,] 0.5725490 0.5686275 0.5647059 0.5686275 0.5686275 0.5607843
## [5,] 0.5725490 0.5647059 0.5490196 0.5647059 0.5686275 0.5647059
## Image
##   colorMode    : Color
##   storage.mode : double
##   dim          : 4608 2592 3
##   frames.total : 3
##   frames.render: 1
##
## imageData(object)[1:5,1:6,1]
```

```
##           [,1]      [,2]      [,3]      [,4]      [,5]      [,6]
## [1,] 0.5098039 0.5254902 0.5450980 0.5411765 0.5333333 0.5333333
## [2,] 0.5215686 0.5215686 0.5333333 0.5450980 0.5372549 0.5333333
## [3,] 0.5215686 0.5176471 0.5333333 0.5450980 0.5411765 0.5372549
## [4,] 0.5137255 0.5098039 0.5215686 0.5411765 0.5450980 0.5450980
## [5,] 0.5098039 0.5058824 0.5137255 0.5254902 0.5450980 0.5529412
```

```
#print(myimages[[1]])
#display(myimages[[1]])
#summary(myimages[[1]])
#hist(myimages[[1]])
#str(myimages[[1]])
```

In the code shown below, we plot an image of each of four diseases.

```
par(mfrow = c(2,2))
for (i in m) {plot(myimages[[i]])}
```

In the R code chunk shown below, the histogram of RBG channels are shown. From the figures, it is clear that the intensity of RBG colors for each of the four disease images are quite different from each other. Intensity values range between 0 and 1.

```
par(mfrow = c(1,1))
for (i in m) {hist(myimages[[i]])}
```

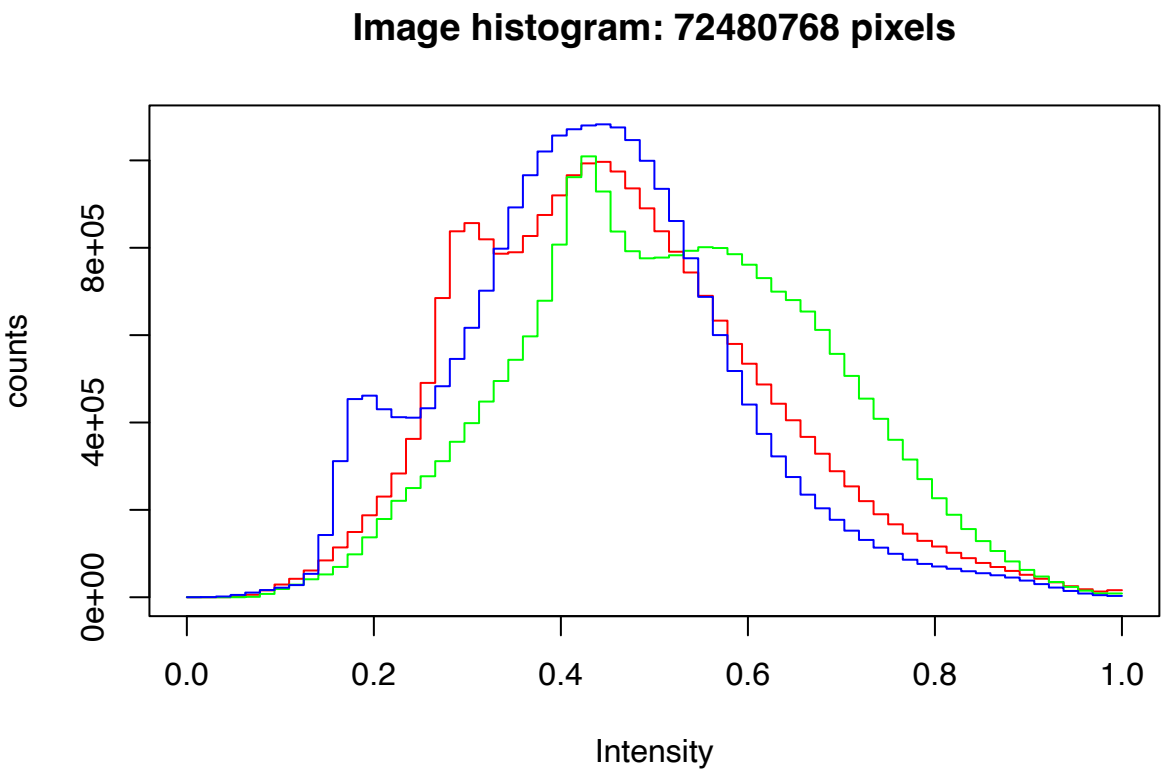


Image histogram: 72480768 pixels
counts
0e+00
4e+05
8e+05
0.0
0.2
0.4
0.6
0.8
1.0
Intensity

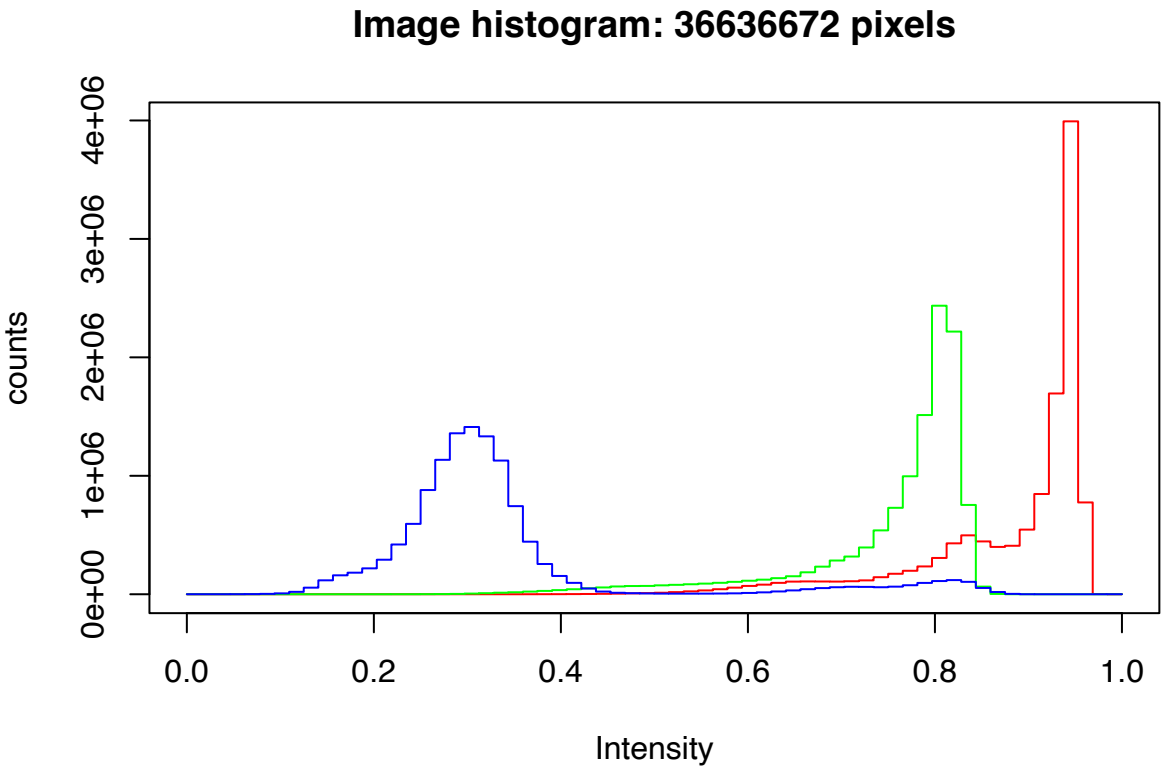


Image histogram: 36636672 pixels
counts
0e+00
1e+06
2e+06
3e+06
4e+06
0.0
0.2
0.4
0.6
0.8
1.0
Intensity

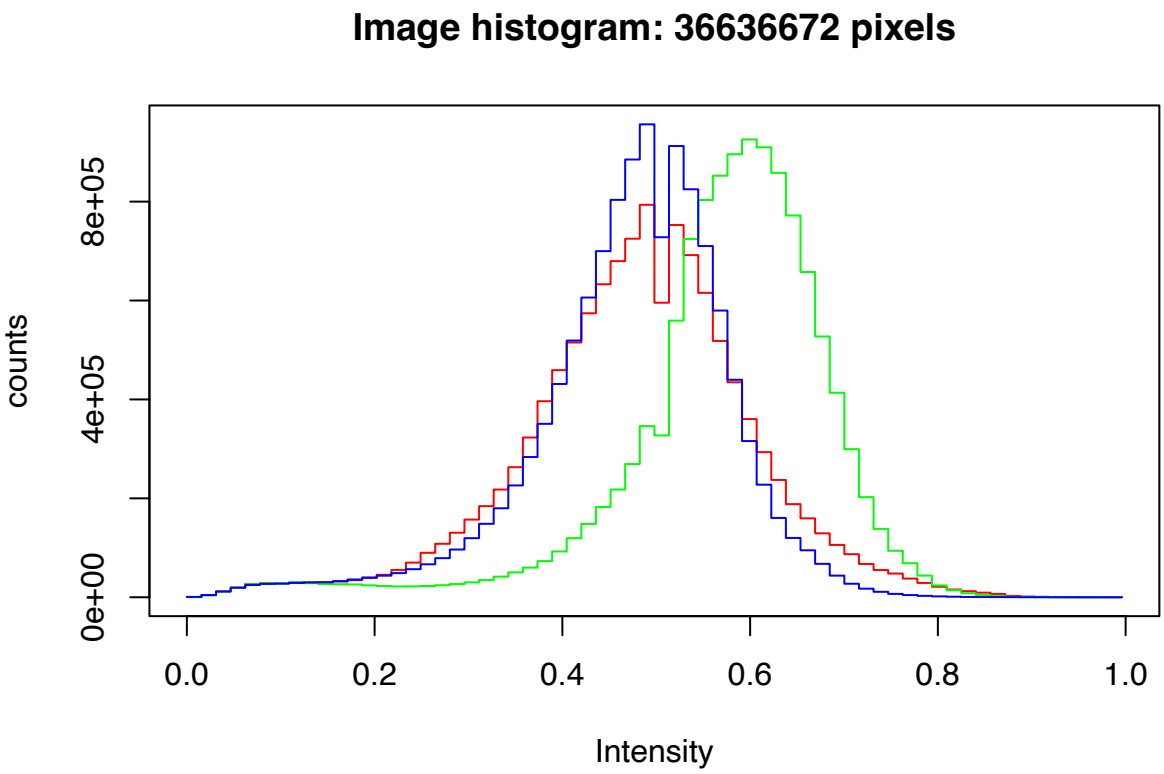
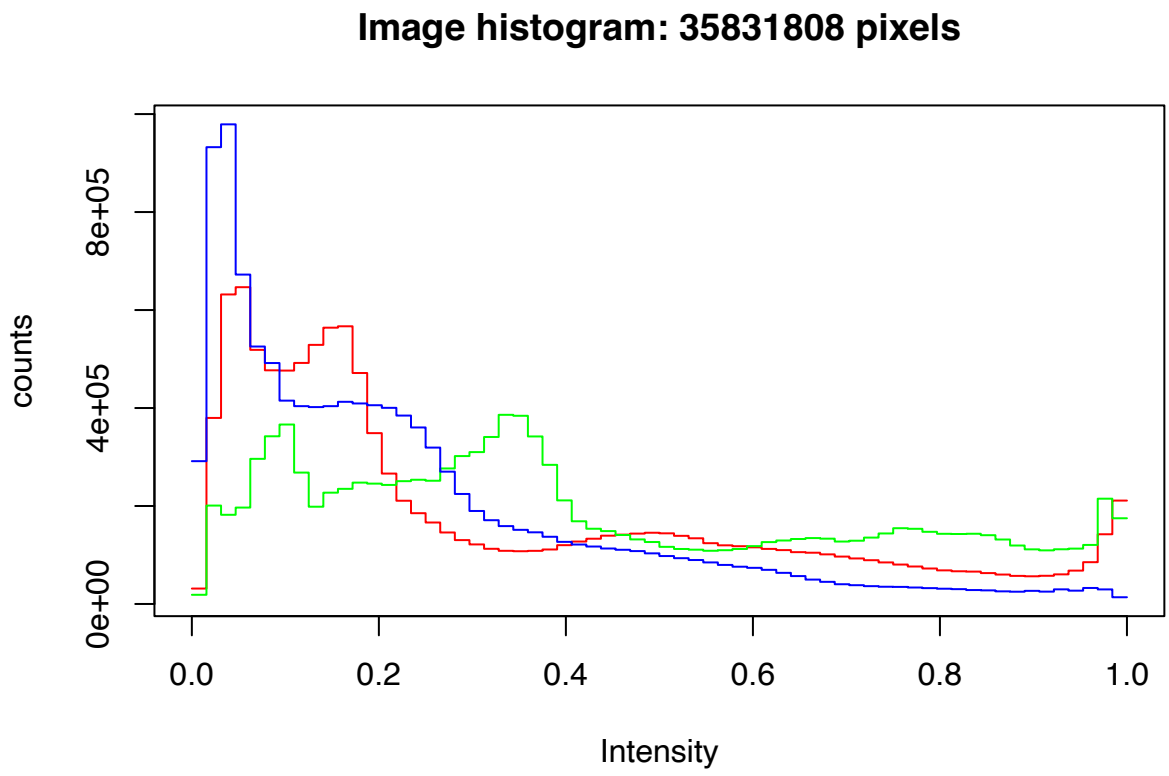


Image histogram: 36636672 pixels
counts
0e+00
4e+05
8e+05
0.0
0.2
0.4
0.6
0.8
1.0
Intensity



Step 4. Resizing and Reshaping Images

As we already know, the size of the images are different. As part of data preparation, one needs to convert all the images into a one fixed size. Here we are converting their size and reshaping into **36 times 36 times 3**.

```
for (i in 1:length(images)) {myimages[[i]] <- resize(myimages[[i]], 36, 36)}
for (i in 1:length(images)) {myimages[[i]] <- array_reshape(myimages[[i]], c(36, 36, 3))}
```

Step 5. Row Binding All the Images into Training and Test Sets

In this step, we bind all the images into rows and divide them into three sets, including training, validation, and test sets. The training set contains the 14 images of each disease. Validation and test sets contain 2 images and 4 images of each disease, respectively.

```
library(tensorflow)
#training set
tr <- c(1:14, 21:34, 41:54, 61:74)
x.train <- NULL
for (i in tr) {x.train <- rbind(x.train, myimages[[i]])}
str(x.train)
```

```
##  num [1:56, 1:3888] 0.482 0.45 0.271 0.259 0.717 ...
```

```
#validation set
va <- c(15:16, 35:36, 55:56, 75:76)
x.valid <- NULL
for (i in va) {x.valid <- rbind(x.valid, myimages[[i]])}

#test set
te <- c(17:20, 37:40, 57:60, 77:80)
x.test <- NULL
for (i in te) {x.test <- rbind(x.test, myimages[[i]])}
```

In the code chunk shown below, we are defining the Y variables for each of the diseases. That is, in the training set, the first 15 images are related to BB disease, while the next 15 images related to BP disease and so on. Similarly, for the test set the first 5 images with BB disease, next five with BP disease and so on.

```
y.train <- c(0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
             2,2,2,2,2,2,2,2,2,2,2,2,2,3,3,3,3,3,3,3,3,3,3,3,3,3,3)
y.valid <- c(0,0,1,1,2,2,3,3)
y.test <- c(0,0,0,0,1,1,1,1,2,2,2,2,3,3,3,3)
```

Step 6. One-Hot Encoding/Binary/Dummy Variables

Here, the binary or dummy variables are defined for each of the four diseases.In machine learning, this is also called one-hot encoding.

```
train.labels <- to_categorical(y.train)
```

```
## Loaded Tensorflow version 2.8.0
```

```
valid.labels <- to_categorical(y.valid)
test.labels <- to_categorical(y.test)

#train.labels
#test.labels
```

Step 7. Model Building

In model building, we start by creating a sequential model and then add various layers. ReLu (Rectified Linear Unit) is used as an activation function for hidden layers. The model is acivated with **36 times 36 times 3** that equal 3888, which inturn goes into the **input_shape**. For the output layer, we use **softmax** as an activation function. Finally, the summary of the model is obtained from the **summary** function.

```
str(x.train)
```

```
##  num [1:56, 1:3888] 0.482 0.45 0.271 0.259 0.717 ...
```

```
model1 <- keras_model_sequential()
model1 %>%
  layer_dense(units = 256, activation = 'relu', input_shape = c(3888)) %>%
  layer_dense(units = 128, activation = 'relu') %>%
  layer_dense(units = 4, activation = 'softmax')

summary(model1)
```

```
## Model: "sequential"
## _________________________________________________________________________
##  Layer (type)                       Output Shape                    Param #
## =========================================================================
##  dense_2 (Dense)                    (None, 256)                     995584
##
##  dense_1 (Dense)                    (None, 128)                     32896
##
##  dense (Dense)                      (None, 4)                       516
##
## =========================================================================
## Total params: 1,028,996
## Trainable params: 1,028,996
## Non-trainable params: 0
## _________________________________________________________________________
```

Step 8. Showing the Calculations of Total Number of Parameters

The R code chunk below explains how we got the total number of parameters as 1028996 in the above step. The number that is added for each of the line below are intercepts.

```
(3888*256)+256
```

```
## [1] 995584
```

```
(128*256)+128
```

```
## [1] 32896
```

```
(128*4)+4
```

```
## [1] 516
```

```
#Total number of parameters = 1028996
```

Step 9. Compile the Model

In this step, we compile the model. The **categorical_crossentropy** is used for loss as we are doing a multi-class classification model. **Adam** is used as an optimizer while the **accuracy** is used as a metric.

```
model1 %>%
  compile(loss = 'categorical_crossentropy',
          optimizer = 'adam',
          metrics = 'accuracy')
```

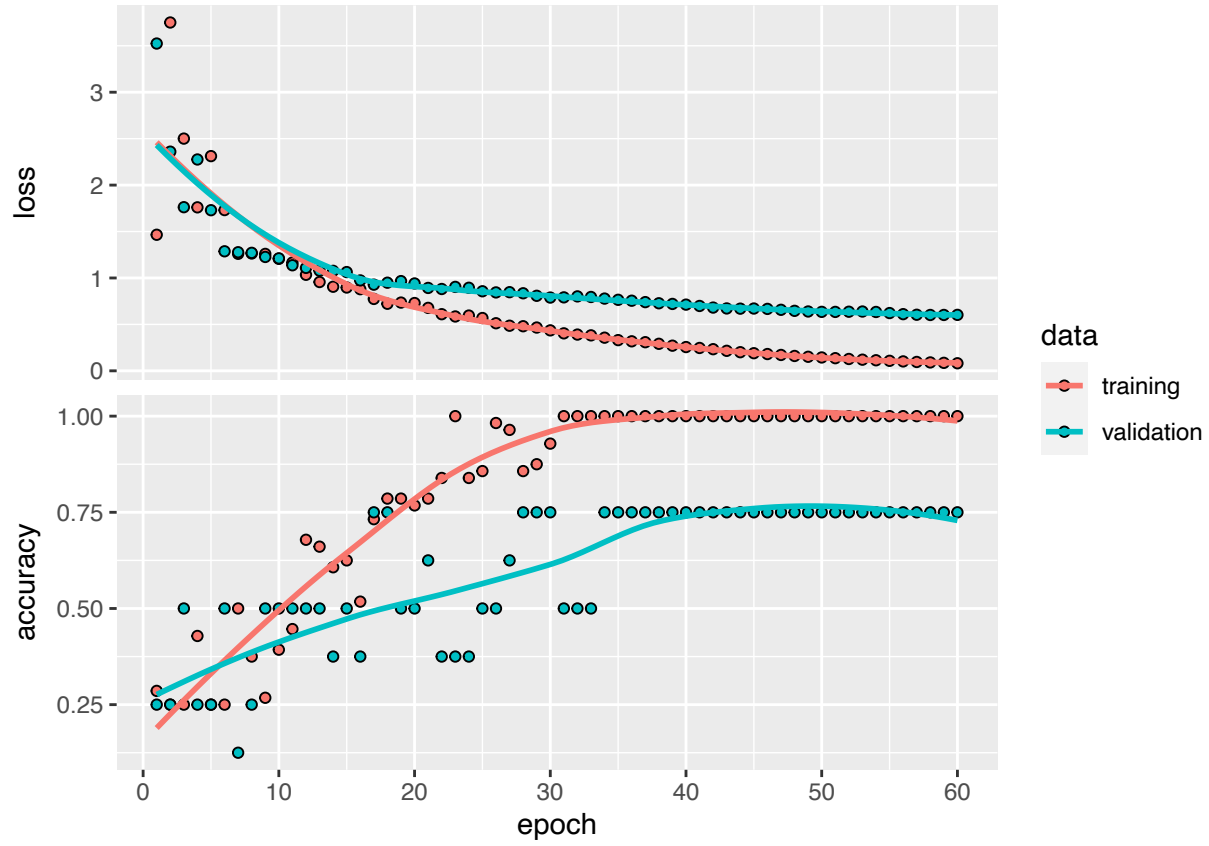
Step 10. Fit the Model

In this step, we fit the model. The plot consists of two panels. The top panel shows loss while the lower panel shows the accuracy for both training and validation sets across the number of epochs, which is shown on the x-axis. From the figure, key takeaway is that at about 22 epochs the accuracy of the classification of disease images remains more or less same for the rest of epochs.

```
history <- model1 %>%
  fit(x.train,
      train.labels,
      epochs = 60,
      batch_size = 65,
      validation_data = list(x.valid, valid.labels))

plot(history)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Step 11. Model Evaluation and Prediction - Train Data

Here, confusion matrix and prediction probabilities of the results on training data is presented.

```
# Model Evaluation and Prediction - Train Data
model1 %>% evaluate(x.train, train.labels)
```

```
##       loss   accuracy
## 0.07698945 1.00000000
```

```r
#confusion matrix
pred <- model1 %>% predict(x.train) %>% k_argmax()
table(Predicted = as.numeric(pred), Actual = y.train)
```

```
##           Actual
## Predicted  0  1  2  3
##         0 14  0  0  0
##         1  0 14  0  0
##         2  0  0 14  0
##         3  0  0  0 14
```

```r
#Prediction probabilities
prob <- model1 %>% predict(x.train)
cbind(round(prob, 3), Predicted_class = as.numeric(pred), Actual = y.train)
```

```
##                                Predicted_class Actual
##  [1,] 0.938 0.001 0.034 0.028               0      0
##  [2,] 0.904 0.003 0.043 0.050               0      0
##  [3,] 0.884 0.004 0.068 0.044               0      0
##  [4,] 0.977 0.000 0.008 0.015               0      0
##  [5,] 0.976 0.001 0.006 0.017               0      0
##  [6,] 0.984 0.000 0.005 0.011               0      0
##  [7,] 0.975 0.000 0.008 0.017               0      0
##  [8,] 0.953 0.000 0.022 0.025               0      0
##  [9,] 0.950 0.001 0.017 0.032               0      0
## [10,] 0.941 0.001 0.019 0.039               0      0
## [11,] 0.874 0.003 0.070 0.053               0      0
## [12,] 0.920 0.001 0.036 0.043               0      0
## [13,] 0.716 0.020 0.158 0.105               0      0
## [14,] 0.809 0.004 0.071 0.116               0      0
## [15,] 0.001 0.987 0.003 0.009               1      1
## [16,] 0.000 0.994 0.002 0.004               1      1
## [17,] 0.000 0.993 0.002 0.004               1      1
## [18,] 0.000 0.991 0.004 0.005               1      1
## [19,] 0.000 0.996 0.002 0.002               1      1
## [20,] 0.000 0.994 0.003 0.003               1      1
## [21,] 0.000 0.993 0.004 0.003               1      1
## [22,] 0.000 0.993 0.002 0.004               1      1
## [23,] 0.002 0.981 0.004 0.013               1      1
## [24,] 0.002 0.971 0.005 0.021               1      1
## [25,] 0.000 0.996 0.001 0.002               1      1
## [26,] 0.002 0.958 0.010 0.031               1      1
## [27,] 0.000 0.990 0.005 0.005               1      1
## [28,] 0.000 0.994 0.002 0.004               1      1
## [29,] 0.081 0.001 0.858 0.059               2      2
## [30,] 0.048 0.004 0.821 0.127               2      2
## [31,] 0.035 0.006 0.882 0.077               2      2
## [32,] 0.030 0.004 0.880 0.085               2      2
## [33,] 0.013 0.001 0.966 0.020               2      2
## [34,] 0.010 0.001 0.970 0.018               2      2
## [35,] 0.009 0.001 0.978 0.012               2      2
## [36,] 0.007 0.001 0.984 0.009               2      2
## [37,] 0.064 0.011 0.822 0.102               2      2
```

```
## [38,] 0.017 0.008 0.948 0.027              2      2
## [39,] 0.026 0.004 0.952 0.017              2      2
## [40,] 0.044 0.002 0.917 0.037              2      2
## [41,] 0.014 0.002 0.967 0.017              2      2
## [42,] 0.035 0.002 0.896 0.067              2      2
## [43,] 0.009 0.002 0.055 0.934              3      3
## [44,] 0.008 0.000 0.010 0.982              3      3
## [45,] 0.006 0.000 0.009 0.985              3      3
## [46,] 0.018 0.000 0.023 0.958              3      3
## [47,] 0.033 0.016 0.114 0.838              3      3
## [48,] 0.017 0.006 0.037 0.940              3      3
## [49,] 0.132 0.002 0.104 0.761              3      3
## [50,] 0.056 0.012 0.087 0.845              3      3
## [51,] 0.032 0.008 0.071 0.889              3      3
## [52,] 0.049 0.014 0.038 0.898              3      3
## [53,] 0.077 0.012 0.050 0.862              3      3
## [54,] 0.027 0.013 0.022 0.938              3      3
## [55,] 0.041 0.015 0.063 0.881              3      3
## [56,] 0.083 0.022 0.087 0.807              3      3
```

Step 12. Evaluation and Prediction on the Test Data

In this final step, the confusion matrix and prediction probabilities of the model evaluated on the test data is presented.

```
# Evaluation and prediction on the test data
model1 %>% evaluate(x.test, test.labels)
```

```
##      loss  accuracy
## 0.7601745 0.6250000
```

```
#confusion matrix
pred <- model1 %>% predict(x.test) %>% k_argmax()
table(Predicted = as.numeric(pred), Actual = y.test)
```

```
##          Actual
## Predicted 0 1 2 3
##         1 0 4 0 0
##         2 4 0 4 2
##         3 0 0 0 2
```

```
#prediction probabilities
prob <- model1 %>% predict(x.test)
cbind(round(prob, 2), Predicted_class = as.numeric(pred), Actual = y.test)
```

```
##                          Predicted_class Actual
##  [1,] 0.13 0.02 0.61 0.24               2      0
##  [2,] 0.17 0.02 0.58 0.23               2      0
##  [3,] 0.15 0.03 0.62 0.21               2      0
##  [4,] 0.13 0.04 0.67 0.16               2      0
##  [5,] 0.00 0.99 0.00 0.00               1      1
```

```
##  [6,] 0.00 0.99 0.00 0.00              1      1
##  [7,] 0.00 0.99 0.00 0.00              1      1
##  [8,] 0.00 0.99 0.00 0.00              1      1
##  [9,] 0.07 0.01 0.77 0.15              2      2
## [10,] 0.06 0.01 0.79 0.15              2      2
## [11,] 0.07 0.00 0.76 0.17              2      2
## [12,] 0.07 0.01 0.82 0.11              2      2
## [13,] 0.08 0.01 0.81 0.10              2      3
## [14,] 0.06 0.17 0.39 0.38              2      3
## [15,] 0.01 0.00 0.04 0.95              3      3
## [16,] 0.01 0.00 0.02 0.97              3      3
```

Step 13. Results, Conclusions, and Future Steps

The training results show good results where the accuracy is quite high but when it comes to the results of test set there are some misclassified disease images. The results show that the analysis suffers from overfitting. It appears that the number and quality of images were not sufficient to generalize and predict the disease outside of the training pool accurate enough.

For future analysis, one needs to focus on collecting the disease images in such as way that it captures disease at various stages of the plant and overall diversity of the disease images were needed for the model to able to generalize. Several other strategies are also available to improve the accuracy in the test set. But, that is for another day.