Assignment 6: Fit a CNN to a HuggingFace Dataset
Keith Tran - 300240028

The dataset chosen for this assignment is the "beans" dataset from HuggingFace. This dataset contains images of bean leaves classified into three categories: Healthy, Angular Leaf Spot, and Bean Rust. The data comes organized into predefined training, validation, and test splits (1034, 133, and 128 samples respectively). Preprocessing involved resizing images to a fixed size of 128x128 pixels and normalizing pixel values to [0,1] scale.

A Sequential model – a linear stack of layers where data flows sequentially from one layer to the next – was used and comprised of:

- An input layer: defines the shape of input data (128x128 pixels), as well as specifying three colour channels (red, blue, green)
- Three convolutional blocks: apply convolution (each increasing in scope), ReLU activation, and MaxPooling for downsampling to focus on dominant features
- Flattening layer: converts 3D feature maps into a 1D vector to be used by dense layers
- Dense layers: a fully connected layer that combines features and applies the ReLU activation function
- Dropout layer: sets 50% of outputs of the previous layer to zero to prevent overfitting
- Output layer: applies a softmax function that converts raw scores into probabilities spread across each of the three output classes

Below is the classification report on the model's performance on the test set of 128 samples. The accuracy decently good at 0.80. Based on recall scores, the model is better at predicting Healthy images and worse at predicting Bean Rust images. According to the confusion matrix, a significant amount of Angular Leaf Spot images are falsely predicted as Bean Rust images. This could indicated that the two image classes are too similar and are confusing the model, or there may be a class imbalance favouring Bean Rust samples in the dataset.

	Precision	Recall	F1-score	Support
Healthy	0.84	0.88	0.86	42
Angular Leaf Spot	0.80	0.81	0.80	43
Bean Rust	0.78	0.72	0.75	43
Accuracy			0.80	128
Macro average	0.80	0.81	0.80	128
Weight Average	0.80	0.80	0.80	128

The training and validation curves for both accuracy and loss are fairly similar. The both steadily increase over epochs, although validation loss begins to increase afer the eightth epoch, suggesting slight overfitting. This could be rectified by increasing dropout, or altering the dataset.