Implementation & Testing – Model Development Phase 5

Following the inability to clarify the contradictory outcomes of the last few tests in phase 4, a comprehensive review of the project was performed, with particular focus given to phases 3 and 4.

Phase 3, which introduced pre-trained architectures to the project, produced outstanding results when utilizing the *MobileNetV2* architecture alongside the *LeafSnap\_15\_Lab* dataset in 3.1, which obtained a test accuracy of 99.6% and a test loss of 0.05, as well as a confident confusion matrix. Later tests of this phase also produced promising results as the *shrooms\_ds* dataset was implemented in 3.2, then balanced and expanded in 3.3, the latter of which obtained a test accuracy of 76.3% and a test loss of 0.68. Test 3.4 then introduced an alternative pre-trained model, *EfficientNetB0*, but unfortunately generated poor test results of 13.3% accuracy and 2.20 loss, attributed to a failure to optimize the code to the new architecture prior to training.

Although the final results of phase 3 weren’t promising, optimism was still held for the use of *EfficientNet* architectures within the project, and so phase 4 was dedicated to this completing this objective. Two variations of the architectural family were implemented, *EfficientNetB0* and *EfficientNetV2S*, and throughout phase 4 the performance of these architectures was compared to determine which was optimal. From the very first test of this phase, 4.1, great potential could be seen in the use of these pre-trained models, as the test results produced were an accuracy of 82.1% and a loss of 0.57. Furthermore, the test results of this phase only improved, as between 4.3 and 4.6 all test accuracies values were at least 99%, and the final test 4.7 produced an accuracy of 94.5% and a loss of 0.18 – truly great results.

That being said, not all the output data was promising, as difficulty was experienced in generating optimistic confusion matrices during the entirety of phase 4. Even in tests where the accuracy and loss values neared perfection, the accompanying confusion matrices told a contradictory and opposing story. Further confirmation of performance issues was found when classification reports were introduced, as they supported the confusion matrices in opposing the narrative of the great accuracy and loss results of the various tests of phase 4. Even when measures were taken to minimise the discrepancy between the performance metrics, such as balancing the datasets, applying class weights, adjusting hyperparameters and returning to a simpler dataset, no substantial progress was made. Although test 4.7 produced the brilliant accuracy and loss results mentioned above, its confusion matrix and classification report failed to align themselves with this narrative of success, remaining in contradiction weighted average values of 0.103, 0.104, and 0.103 for precision, recall and F1-Score respectively.

When broadly comparing the results of phases 3 and 4, it is clear that great accuracy and loss values can be found throughout both. However, it must be stated that phase 3 displays more immediate potential, as its confusion matrices are far more congruent with their respective accuracies and losses than those of phase 4. While substantial efforts were made to solve the issue of incongruency within the latter phase, progress was negligible.

Therefore, in conclusion of this review, the decision has been made to continue the model development process based on phase 3 instead of phase 4. Ultimately, this means abandoning the *EfficientNet* family of architectures in favour of *MobileNetV2*, and utilising the code of phase 3 as the foundation on which the project’s final outcome can be reached.

In addition to the regression to an earlier code for model development, phase 5 will utilise a new dataset, *17flowers*, which contains 17 classes of different flower species. This decision was made based upon the fact that the dataset of leaves has already been discarded due to its lab-based images being too pristine, in addition to the fact that the mushroom dataset has not produced results which combine fantastic accuracy and congruent supporting performance metrics. These issues left a window of opportunity for this new dataset, which features realistic, real-world images similar to the mushroom dataset, but will hopefully produce better and more congruent outcomes.

5.1

As stated above, the code responsible for model development during phase 5 is actually based upon that used during phase 3.While there is a lot of fundamental overlap between the codes used during phases 3 and 4, the biggest differences are the exclusion of the classification report, which was only introduced during phase 4, as well as the use of a different pre-trained architecture, as phase 3 utilised a variety of the *MobileNet* architecture, while phase 4 explored the *EfficientNet* family. At this current moment, the plan is to stick to *MobileNetV2*, the architecture which demonstrated great promise during the tests in which it was implemented. However, the additional performance metrics from phase 4 will be introduced during this current phase, as they are crucial tools in determining the efficacy of the models.

5.1 will consist of a series of tests, in order to allow for the optimal configuration of the hyperparameter values. The first test within this series, 5.1a, is discussed below.

5.1a

Since test 3.1 produced the best results of its phase, its code was chosen as the base for developing the model in test 5.1. The hyperparameter values remained identical, as the values used were a learning rate of 0.0001, a batch size of 32, and an epoch value of 10. The only real difference between 3.1 and 5.1 is the dataset used for training, as the *Leafsnap\_15\_Lab* is replaced by *flowers17*.

A black screen with white text

AI-generated content may be incorrect.

As can be seen in the training results below, this initial version of phase 5 appears to have generated highly promising results. Training accuracy rises continually, from 25.5% initially to 93.5% by the final epoch, while the training loss improves from 2.54 to 0.27. The validation results provide further confirmation of the training process with the final few epochs reaching an average accuracy of around 94% and a final validation loss value of 0.24.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Epoch | accuracy | loss | val\_accuracy | val\_loss |
| 1 | 0.2552083432674408 | 2.539233684539795 | 0.65234375 | 1.3769510984420776 |
| 2 | 0.581250011920929 | 1.352041482925415 | 0.8359375 | 0.7845145463943481 |
| 3 | 0.715624988079071 | 0.9379004836082458 | 0.8671875 | 0.6117345690727234 |
| 4 | 0.8166666626930237 | 0.6635771989822388 | 0.91015625 | 0.49095413088798523 |
| 5 | 0.8374999761581421 | 0.5551220774650574 | 0.9140625 | 0.4199945032596588 |
| 6 | 0.8552083373069763 | 0.4924551546573639 | 0.92578125 | 0.33591848611831665 |
| 7 | 0.8999999761581421 | 0.38867369294166565 | 0.921875 | 0.34080207347869873 |
| 8 | 0.90625 | 0.3629199266433716 | 0.9375 | 0.31862807273864746 |
| 9 | 0.8989583253860474 | 0.35086092352867126 | 0.9453125 | 0.25754785537719727 |
| 10 | 0.9354166388511658 | 0.2749722898006439 | 0.94140625 | 0.23866546154022217 |

The success of the training stage is reinforced by equally fantastic test results, as an accuracy of 96.5% and a loss of 0.24 were achieved.

|  |  |
| --- | --- |
| Test Loss | 0.2416735291481018 |
| Test Accuracy | 0.9652777910232544 |

The close proximity of the training, validation and test result values suggest a confident and successful learning process, with minimal indication of overfitting. The graphs below display the accuracy and loss values for both the training and validation stages, and provide a visual demonstration of the success of these stages. One point worth noting here is that the plots of both the accuracy and loss graphs do not plateau within the allotted number of epochs. This suggests that increasing this value could allow the model to increase its performance further. This will be taken into account during the next test.

A graph of the same number

AI-generated content may be incorrect.

Finally, the confusion matrix for 5.1a validates its test results, as the desirable diagonal line of true positives is prominent, suggesting that only a handful of predicted labels were incorrect. Given the extreme difficulty experienced in attempting to align confusion matrices with its accompanying results during phase 4, a supportive confusion matrix such as this one for 5.1a is a welcome sight.

A chart of different colored squares

AI-generated content may be incorrect.

5.1b

Based on the result of the 5.1a, in which the accuracy and loss plots did not yet reach a plateau, this next test aims to explore the possibility of achieving better results by simply increasing the number of epochs. In 5.1b, this value will be increased from 10 to 15, while all other hyperparameters and variables will be kept the same.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Epoch | accuracy | loss | val\_accuracy | val\_loss |
| 1 | 0.22499999403953552 | 2.605189085006714 | 0.66015625 | 1.387690782546997 |
| 2 | 0.5874999761581421 | 1.3761347532272339 | 0.84765625 | 0.8179001212120056 |
| 3 | 0.7447916865348816 | 0.8805817365646362 | 0.86328125 | 0.6340733766555786 |
| 4 | 0.8072916865348816 | 0.68577641248703 | 0.90625 | 0.4802386462688446 |
| 5 | 0.8656250238418579 | 0.536548912525177 | 0.91015625 | 0.4210643768310547 |
| 6 | 0.875 | 0.46937209367752075 | 0.94921875 | 0.33501434326171875 |
| 7 | 0.9041666388511658 | 0.39783230423927307 | 0.9375 | 0.3226008415222168 |
| 8 | 0.9020833373069763 | 0.36271336674690247 | 0.9375 | 0.30015283823013306 |
| 9 | 0.9208333492279053 | 0.31864750385284424 | 0.953125 | 0.255307674407959 |
| 10 | 0.9437500238418579 | 0.26489657163619995 | 0.9453125 | 0.2481955885887146 |
| 11 | 0.9312499761581421 | 0.2660682499408722 | 0.9609375 | 0.21062098443508148 |
| 12 | 0.9437500238418579 | 0.2507191002368927 | 0.97265625 | 0.19044290482997894 |
| 13 | 0.9572916626930237 | 0.21686074137687683 | 0.96875 | 0.1858711689710617 |
| 14 | 0.953125 | 0.19487789273262024 | 0.97265625 | 0.17034241557121277 |
| 15 | 0.9677083492279053 | 0.1592138260602951 | 0.97265625 | 0.14297834038734436 |

As can be seen in the training and validation results table above, it appears that the 5 additional epochs provided the model with the time to improve both its accuracy and loss values beyond those achieved during 5.1a. The training and validation accuracy values average around 97%, an improvement of around 3%, while the average value of the training and validation loss values is around 0.15, showing an improvement of at least 0.10 in comparison to 5.1a.

|  |  |
| --- | --- |
| Test Loss | 0.26249635219573975 |
| Test Accuracy | 0.9375 |

Unfortunately, this improvement did not permeate its way into the test results, as for some reason the test results of 5.1b are actually worse than those of 5.1a, even though the training and validation data showed promise for improvement. The test loss of 5.1b is 0.26, which is only 0.02 worse than the equivalent value of 5.1a, however the test accuracy of 5.1b is 93.8%, which is around 3% worse than its predecessor. Initially it was speculated that the test results of 5.1a and 5.1b had been mislabelled, but after verifying that no error was made regarding results storage, the drop in performance between training and testing is simply confusing. One possible explanation would be that the images designated to the test stage were more challenging for the model in 5.1b than in 5.1a. Aside from this potential reason, it is not clear at this point what else could cause the incongruency between the training & validation and the test results.

A graph of different colored lines

AI-generated content may be incorrect.

In more optimistic news, the accuracy and loss plots from the training and validation stages suggest less overfitting than those of 5.1a, as there is less distance between the training and validation plots while superior values are reached. Additionally, even though the number of epochs was extended from 5.1a to 5.1b, the plots fail to demonstrate plateauing, suggesting that results could again be improved further by simply increasing the number of epochs.

A chart with different colored squares

AI-generated content may be incorrect.

Interestingly, although the test results of 5.1b were inexplicably worse than 5.1a, the improvement seen in its training and validation stages is reflected in the confusion matrix which only contains 6 incorrect predictions, marginally better than the 9 incorrect labels of 5.1a’s confusion matrix.

5.1c

As was stated in the analysis of 5.1b, its accuracy and loss plots do not reach a point of plateau before completing the final epoch, suggesting that increasing this value further could give the model the opportunity to produce even better results. Therefore, 5.1c will include an additional 5 epochs for its training stage, making the number of epochs 20. Additionally, to ensure a more gradual learning process, the learning rate has been decreased to 0.00005.

The table below contains the accuracy and loss results of the training and validation stages for 5.1c. Interestingly, although the learning rate was halved between tests, the rate of improvement for all metrics in 5.1c is only marginally less than the 5.1b. Throughout the entire table, the values obtained for each epoch are almost identical in comparison to 5.1b.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Epoch | accuracy | loss | val\_accuracy | val\_loss |
| 1 | 0.23229166865348816 | 2.591217041015625 | 0.6796875 | 1.453842043876648 |
| 2 | 0.5885416865348816 | 1.3670449256896973 | 0.80078125 | 0.8213101029396057 |
| 3 | 0.7489583492279053 | 0.8703815340995789 | 0.875 | 0.6153343915939331 |
| 4 | 0.7906249761581421 | 0.6898549199104309 | 0.90625 | 0.4811309576034546 |
| 5 | 0.8614583611488342 | 0.5412112474441528 | 0.94140625 | 0.4060458540916443 |
| 6 | 0.8697916865348816 | 0.46641620993614197 | 0.94140625 | 0.34133467078208923 |
| 7 | 0.8854166865348816 | 0.38851770758628845 | 0.9375 | 0.3291257321834564 |
| 8 | 0.90625 | 0.34926652908325195 | 0.9453125 | 0.29308176040649414 |
| 9 | 0.9260416626930237 | 0.30114060640335083 | 0.953125 | 0.2475132793188095 |
| 10 | 0.9458333253860474 | 0.2504172623157501 | 0.953125 | 0.2465241402387619 |
| 11 | 0.9375 | 0.2651033401489258 | 0.97265625 | 0.20486091077327728 |
| 12 | 0.9364583492279053 | 0.23950694501399994 | 0.97265625 | 0.17259147763252258 |
| 13 | 0.9583333134651184 | 0.19945861399173737 | 0.97265625 | 0.1770787537097931 |
| 14 | 0.9593750238418579 | 0.18303419649600983 | 0.97265625 | 0.1593366414308548 |
| 15 | 0.9635416865348816 | 0.1842469722032547 | 0.984375 | 0.13526296615600586 |
| 16 | 0.9624999761581421 | 0.1698601096868515 | 0.97265625 | 0.16807278990745544 |
| 17 | 0.9624999761581421 | 0.16184809803962708 | 0.9765625 | 0.13906119763851166 |
| 18 | 0.9708333611488342 | 0.140947625041008 | 0.96484375 | 0.14675182104110718 |
| 19 | 0.965624988079071 | 0.14761771261692047 | 0.97265625 | 0.10969216376543045 |
| 20 | 0.96875 | 0.13121174275875092 | 0.9609375 | 0.145465686917305 |

Another interesting point is that increasing the number of epochs from 15 to 20 did not have substantial positive impact on the model’s performance during the training and validation stages: slight improvement continues to occur regarding the training accuracy and loss, however the best validation accuracy value, 98.4%, was produced during epoch 15. As for validation loss, the best value of 0.11 was produced during epoch 19.

Although the training and validation data does not seem to have improved substantially, the test results of 5.1c provide confirmation that these new hyperparameter values had a positive impact on the model’s performance, as a test accuracy of 97.2% and a test loss of 0.14 were recorded. These are actually the best test results of phase 5 so far.

|  |  |
| --- | --- |
| Test Loss | 0.14287321269512177 |
| Test Accuracy | 0.9722222089767456 |

The accuracy and loss graphs visually explain the outcome of 5.1c, and the possible benefits of its hyperparameters: firstly, the increased number of epochs allows the plots to finally display plateauing, while the reduced learning rate could help explain the closer proximity of the training and validation plots, suggesting an improved learning process.

A graph of different colored lines

AI-generated content may be incorrect.

Once again, this test has produced a very promising confusion matrix, with only 7 incorrect predictions. This in line with the other tests of phase 5, and in contrast to the very confusing confusion matrices of phase 4.

A colorful squares with white text

AI-generated content may be incorrect.

5.1d

Now that minor adjustments to the number of epochs and learning rate have been explored, 5.1d will investigate the impact of changing the batch size, a hyperparameter which has been kept consistent to this point. The new value for the batch size will be 16, which is half the previous value of 32. The decision to reduce the batch size was made based on the fact that the dataset in use, flowers17, is relatively small, and so it seemed logical to explore the use of a smaller batch size, as opposed to a larger one. The learning rate and number of epochs are 0.00005 and 20 respectively.

The table below contains the accuracy and loss results from the training and validation stages. It is interesting to note that this is the first time in the recent history of this project that the validation stages were outperformed by their corresponding training stages: typically, the validation accuracy and loss results are superior, however it is the training values which are better in 5.1d. From around the 10th epoch onwards the training data produces the better results, and this dynamic remains until the end, with accuracy and loss values which are respectively 3-5% and 0.06-0.07 better than their validation equivalents throughout.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Epoch | accuracy | loss | val\_accuracy | val\_loss |
| 1 | 0.2955508530139923 | 2.3196723461151123 | 0.7573529481887817 | 1.0593253374099731 |
| 2 | 0.6832627058029175 | 1.0622674226760864 | 0.845588207244873 | 0.667885959148407 |
| 3 | 0.804025411605835 | 0.6844117641448975 | 0.8492646813392639 | 0.5305262207984924 |
| 4 | 0.8548728823661804 | 0.5298187732696533 | 0.9117646813392639 | 0.4303551912307739 |
| 5 | 0.8887711763381958 | 0.41587138175964355 | 0.9117646813392639 | 0.344769686460495 |
| 6 | 0.8961864113807678 | 0.3674969971179962 | 0.9154411554336548 | 0.30353501439094543 |
| 7 | 0.9247881174087524 | 0.28137490153312683 | 0.9411764740943909 | 0.28230857849121094 |
| 8 | 0.9470338821411133 | 0.24932172894477844 | 0.9558823704719543 | 0.2555859088897705 |
| 9 | 0.9523305296897888 | 0.22340647876262665 | 0.9522058963775635 | 0.22667516767978668 |
| 10 | 0.9629237055778503 | 0.17945550382137299 | 0.9375 | 0.23072125017642975 |
| 11 | 0.9724576473236084 | 0.1548006683588028 | 0.9338235259056091 | 0.26181286573410034 |
| 12 | 0.9671609997749329 | 0.14613685011863708 | 0.9411764740943909 | 0.22436897456645966 |
| 13 | 0.9735169410705566 | 0.13485397398471832 | 0.966911792755127 | 0.1842106580734253 |
| 14 | 0.9724576473236084 | 0.12216592580080032 | 0.9595588445663452 | 0.16538400948047638 |
| 15 | 0.9777542352676392 | 0.1146000325679779 | 0.9522058963775635 | 0.16424496471881866 |
| 16 | 0.9872881174087524 | 0.09793656319379807 | 0.9485294222831726 | 0.18462693691253662 |
| 17 | 0.9841101765632629 | 0.09279216080904007 | 0.9485294222831726 | 0.16507244110107422 |
| 18 | 0.9819915294647217 | 0.09327848255634308 | 0.966911792755127 | 0.15214592218399048 |
| 19 | 0.9809321761131287 | 0.0882859155535698 | 0.9595588445663452 | 0.14459627866744995 |
| 20 | 0.9872881174087524 | 0.08013921231031418 | 0.9595588445663452 | 0.15599079430103302 |

Regardless, all values obtained by the final epoch are very promising: 98.7% and 0.08 for the training accuracy and loss, and 96% and 0.15 for validation. Those training results are the best scores obtained in phase 5 so far. However, when comparing 5.1d to its predecessor, it can be seen that 5.1d’s test accuracy of 96.5% is roughly 0.7% worse than its predecessor, and its test loss of 0.22 is worse by around 0.08.

|  |  |
| --- | --- |
| Test Loss | 0.22222426533699036 |
| Test Accuracy | 0.9652777910232544 |

This interesting combination of superior training and inferior validation and accuracy results is indicative of the fact that the model has overtrained on the training set. This point is reinforced by the separation between the training and validation plots in both the accuracy and validation graphs below.

A graph of different colored lines

AI-generated content may be incorrect.

Even though this overfitting during training has occurred, these results are still very useful and informative. This conclusion is supported by the fact that the confusion matrix produced contains only 8 incorrect predictions. That being said, 5.1d did not overtake its predecessors in terms of performance, and so the batch size moving forward will return to 32.

A colorful squares with white text

AI-generated content may be incorrect.

5.2

Now that optimal hyperparameter values have been identified, the next experimentation stage will explore the effect of increasing the input dataset on the performance metrics.

In order to increase the volume of the dataset, a small python program was created which accepts the input dataset as well as a specified factor of multiplication, and applies TensorFlow augmentation methods to the images within each class until they have all been increased to the desired volume. For 5.2, the specified factor of multiplication was 5, and so the new dataset used here has been named *flowers17\_5x*.

INSERT IMAGE OF ORIGINAL MULTIPLIER CODE.

COVER THE HYPERPARAMETER VALUES USED

mobilenet\_LR5e-05\_BS16\_E20

The results table for 5.2 below contains very promising data: the performance metrics obtained during epoch 20 are the best of phase 5 to this point, with an average accuracy of 97.6% and an average loss of 0.073.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Epoch | accuracy | loss | val\_accuracy | val\_loss |
| 1 | 0.6262626051902771 | 1.2390668392181396 | 0.8948529362678528 | 0.4237695038318634 |
| 2 | 0.8707912564277649 | 0.45019522309303284 | 0.9426470398902893 | 0.2414090484380722 |
| 3 | 0.9088804721832275 | 0.3033820688724518 | 0.9544117450714111 | 0.20112556219100952 |
| 4 | 0.9307659864425659 | 0.23659692704677582 | 0.9624999761581421 | 0.1620754897594452 |
| 5 | 0.940025269985199 | 0.20240691304206848 | 0.9602941274642944 | 0.14847072958946228 |
| 6 | 0.9467592835426331 | 0.1736537665128708 | 0.9676470756530762 | 0.12603363394737244 |
| 7 | 0.9564393758773804 | 0.14792826771736145 | 0.9750000238418579 | 0.11362481117248535 |
| 8 | 0.9593855142593384 | 0.13433901965618134 | 0.9816176295280457 | 0.0986107885837555 |
| 9 | 0.9600168466567993 | 0.13438545167446136 | 0.9676470756530762 | 0.11950311064720154 |
| 10 | 0.9595959782600403 | 0.12929010391235352 | 0.970588207244873 | 0.10345669090747833 |
| 11 | 0.9646464586257935 | 0.11466462165117264 | 0.9808823466300964 | 0.08283430337905884 |
| 12 | 0.9640151262283325 | 0.1116393506526947 | 0.9779411554336548 | 0.0874519795179367 |
| 13 | 0.9696969985961914 | 0.09937536716461182 | 0.9786764979362488 | 0.08043872565031052 |
| 14 | 0.9661195278167725 | 0.10111825913190842 | 0.9742646813392639 | 0.0850747600197792 |
| 15 | 0.9739057421684265 | 0.09111125767230988 | 0.979411780834198 | 0.07675392180681229 |
| 16 | 0.9711700081825256 | 0.09318546950817108 | 0.9823529124259949 | 0.07396939396858215 |
| 17 | 0.9675925970077515 | 0.0929829329252243 | 0.9764705896377563 | 0.08187142014503479 |
| 18 | 0.9730639457702637 | 0.08098549395799637 | 0.9808823466300964 | 0.0706314742565155 |
| 19 | 0.9753788113594055 | 0.08018365502357483 | 0.9779411554336548 | 0.08445794135332108 |
| 20 | 0.9747474789619446 | 0.07569650560617447 | 0.9772058725357056 | 0.07234415411949158 |

The test results echo those obtained during training and validation, and again are the best test results obtained during phase 5, with the test accuracy being 98.1% and the test loss being 0.08.

|  |  |
| --- | --- |
| Test Loss | 0.08028818666934967 |
| Test Accuracy | 0.9810495376586914 |

Validation of the tremendous results above is also found in the accuracy and loss plots from the training and validation stages. Both are great, however the loss graph is truly outstanding, as the distance between the training and validation plots is miniscule.

A graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of

AI-generated content may be incorrect.

The confusion matrix also reinforces the success of utilising the enlarged dataset, as even though the volume of the input dataset was 5 times larger than in previous phase 5 stages, only 14 predictions were incorrect, which as a percentage of the total test images is actually far better than preceding tests.

A chart with different colored squares

AI-generated content may be incorrect.

5.3

In order to validate the outstanding results achieved in 5.2, this next stage will re-introduce the classification report. Ideally, the report will confirm that this current configuration is as great as is implied by the performance metrics currently in use.

In order to remain consistent and comparable with the previous stage, 5.3 will include no changes apart from the re-introduction of the classification report.

mobilenet\_LR5e-05\_BS16\_E20

The table below contains the accuracy and loss data from the training and validation stages of 5.3. As seen in 5.2, this current configuration of dataset and hyperparameters has again produced amazing results which are almost identical to those of the previous stage. This was of course expected, since no model-related details were changed.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Epoch | accuracy | loss | val\_accuracy | val\_loss |
| 1 | 0.6256313323974609 | 1.2641345262527466 | 0.9058823585510254 | 0.42930635809898376 |
| 2 | 0.8619528412818909 | 0.44874125719070435 | 0.9382352828979492 | 0.2549041211605072 |
| 3 | 0.9120370149612427 | 0.30498579144477844 | 0.9551470875740051 | 0.1979990154504776 |
| 4 | 0.9265572428703308 | 0.23819653689861298 | 0.9558823704719543 | 0.16820833086967468 |
| 5 | 0.9478114247322083 | 0.19067534804344177 | 0.966911792755127 | 0.1382799744606018 |
| 6 | 0.9486532211303711 | 0.17341232299804688 | 0.9624999761581421 | 0.12811554968357086 |
| 7 | 0.9473905563354492 | 0.15539051592350006 | 0.9727941155433655 | 0.11250313371419907 |
| 8 | 0.9604377150535583 | 0.1319790780544281 | 0.9727941155433655 | 0.1027107685804367 |
| 9 | 0.9631733894348145 | 0.12396258115768433 | 0.9683823585510254 | 0.11364690959453583 |
| 10 | 0.9654881954193115 | 0.11370489001274109 | 0.9676470756530762 | 0.10800516605377197 |
| 11 | 0.9640151262283325 | 0.11889924854040146 | 0.9808823466300964 | 0.08710624277591705 |
| 12 | 0.9654881954193115 | 0.10811132192611694 | 0.9757353067398071 | 0.09001830965280533 |
| 13 | 0.9667508602142334 | 0.09977269172668457 | 0.979411780834198 | 0.08148282021284103 |
| 14 | 0.9707491397857666 | 0.09966247528791428 | 0.9757353067398071 | 0.08560279756784439 |
| 15 | 0.9680134654045105 | 0.09781938046216965 | 0.9742646813392639 | 0.07973917573690414 |
| 16 | 0.9678030014038086 | 0.09409473836421967 | 0.9750000238418579 | 0.08852533251047134 |
| 17 | 0.9726430773735046 | 0.08419913798570633 | 0.9779411554336548 | 0.08499638736248016 |
| 18 | 0.9705387353897095 | 0.09282692521810532 | 0.9735293984413147 | 0.08458393067121506 |
| 19 | 0.9732744097709656 | 0.08357356488704681 | 0.9772058725357056 | 0.0758606418967247 |
| 20 | 0.9753788113594055 | 0.07606006413698196 | 0.9786764979362488 | 0.08495640009641647 |

The test results achieved are also almost identical, with the exact same test accuracy of 98.1%, and a fractionally better test loss of 0.07.

|  |  |
| --- | --- |
| Test Loss | 0.07193555682897568 |
| Test Accuracy | 0.9810495376586914 |

The accuracy and loss plots demonstrate the success of the training stage, with the close proximity of training and validation plots suggesting that overfitting is not occurring.

A comparison of a graph

AI-generated content may be incorrect.

The confusion matrix of 5.3 is equally as brilliant as 5.2, as they both achieved the exact same percentage of true positives, and only mislabelled 14 test images.

A chart with different colored squares

AI-generated content may be incorrect.

Now for the true purpose of 5.3: examining if the outstanding results above are reflected in the classification report. Given the brilliance of the confusion matrix, it was expected that the classification report would support the optimistic evaluation of the other metrics, and fortunately that is the case. All values within the report are north of 0.90, and the macro and weighted averages at the bottom of the report are over 0.97, meaning that the performance of the model is evidenced in the precision, recall and F1-score values. This classification report reassuringly reflects not only the confusion matrix, but also the overall success of the model development process.

In stark contrast to phase 4, all performance metrics are aligned and unified in providing an optimistic analysis of the models developed during phase 5. Given that near perfect scores have been recorded already, the room for improvement is small, however there are a couple more adjustments to test before concluding the model development process.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class | Precision | Recall | F1-Score | Support |
| Bluebell | 0.90625 | 0.9666666666666667 | 0.9354838709677419 | 30.0 |
| Buttercup | 0.9743589743589743 | 1.0 | 0.987012987012987 | 38.0 |
| Coltsfoot | 1.0 | 0.9555555555555556 | 0.9772727272727273 | 45.0 |
| Cowslip | 0.9666666666666667 | 0.9666666666666667 | 0.9666666666666667 | 30.0 |
| Crocus | 0.9333333333333333 | 1.0 | 0.9655172413793104 | 42.0 |
| Daffodil | 1.0 | 0.90625 | 0.9508196721311475 | 32.0 |
| Daisy | 1.0 | 1.0 | 1.0 | 44.0 |
| Dandelion | 0.9615384615384616 | 1.0 | 0.9803921568627451 | 50.0 |
| Fritillary | 1.0 | 1.0 | 1.0 | 48.0 |
| Iris | 1.0 | 1.0 | 1.0 | 35.0 |
| Lilyvalley | 1.0 | 0.9210526315789473 | 0.958904109589041 | 38.0 |
| Pansy | 1.0 | 0.9523809523809523 | 0.975609756097561 | 42.0 |
| Snowdrop | 0.9487179487179487 | 0.9736842105263158 | 0.961038961038961 | 38.0 |
| Sunflower | 1.0 | 1.0 | 1.0 | 45.0 |
| Tigerlily | 1.0 | 1.0 | 1.0 | 44.0 |
| Tulip | 0.9512195121951219 | 0.975 | 0.9629629629629629 | 40.0 |
| Windflower | 1.0 | 1.0 | 1.0 | 45.0 |
| macro avg | 0.9789461704006182 | 0.9774856872573591 | 0.9777459477636382 | 686.0 |
| weighted avg | 0.9804520796525525 | 0.9795918367346939 | 0.9795886425809673 | 686.0 |

5.4

As the model development process nears its conclusion, there are firstly a handful of adjustments to make in hopes of further increasing model performance. One such adjustment, explored in 5.4, is to reduce the degree of augmentation applied to the dataset during expansion.

LINE OR TWO ABOUT THE DIFFERENCE IN AUGMENTATION BETWEEN 5.4 AND 5.3 (5.6 AND 5.5)

SCREENSHOT OF NEW VALUES

mobilenet\_LR5e-05\_BS16\_E20 CHANGE

The augmentation values applied during previous stages were rather substantial, meaning that the images produced may push the recognition ability of the model too far, causing confusion between classes instead of simply generating additional data for each class. By reducing the degrees of augmentation, the images produced for the enlarged dataset will be of more utility to the model, allowing it to gain a deeper understanding of each class, while hopefully avoiding the confusion caused by acute augmentation.

By examining the results table of 5.4 below, it appears as though reducing the augmentation has had the desired effect, as new high scores were obtained across all columns: 98.4%, 0.054, 99.5% and 0.036 for training accuracy, training loss, validation accuracy and validation loss respectively.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Epoch | accuracy | loss | val\_accuracy | val\_loss |
| 1 | 0.6597222089767456 | 1.130545735359192 | 0.9213235378265381 | 0.35696738958358765 |
| 2 | 0.8992003202438354 | 0.36091750860214233 | 0.9654411673545837 | 0.18275301158428192 |
| 3 | 0.9318181872367859 | 0.24568496644496918 | 0.9735293984413147 | 0.14691409468650818 |
| 4 | 0.9553872346878052 | 0.17170915007591248 | 0.9801470637321472 | 0.10592234879732132 |
| 5 | 0.9583333134651184 | 0.1498570442199707 | 0.9860293865203857 | 0.0874500498175621 |
| 6 | 0.9690656661987305 | 0.1207321509718895 | 0.9852941036224365 | 0.07984044402837753 |
| 7 | 0.9764309525489807 | 0.10084906220436096 | 0.9867647290229797 | 0.06459816545248032 |
| 8 | 0.9762205481529236 | 0.09062345325946808 | 0.9882352948188782 | 0.056459397077560425 |
| 9 | 0.9730639457702637 | 0.09119988232851028 | 0.9867647290229797 | 0.05287778377532959 |
| 10 | 0.9785353541374207 | 0.07676930725574493 | 0.9897058606147766 | 0.05467388406395912 |
| 11 | 0.9764309525489807 | 0.07573401927947998 | 0.9926470518112183 | 0.03934750705957413 |
| 12 | 0.9842171669006348 | 0.06555674225091934 | 0.9889705777168274 | 0.04741937294602394 |
| 13 | 0.9774831533432007 | 0.07023613899946213 | 0.987500011920929 | 0.03992653638124466 |
| 14 | 0.9819023609161377 | 0.06439781188964844 | 0.9904412031173706 | 0.04027123004198074 |
| 15 | 0.9823232293128967 | 0.05667911842465401 | 0.9911764860153198 | 0.03640379011631012 |
| 16 | 0.9852693676948547 | 0.05336226522922516 | 0.9904412031173706 | 0.04627865552902222 |
| 17 | 0.9859007000923157 | 0.05362839251756668 | 0.9941176176071167 | 0.029119273647665977 |
| 18 | 0.9810606241226196 | 0.051260173320770264 | 0.9933823347091675 | 0.028798824176192284 |
| 19 | 0.9842171669006348 | 0.05163523182272911 | 0.987500011920929 | 0.04245174303650856 |
| 20 | 0.9835858345031738 | 0.053631141781806946 | 0.9948529601097107 | 0.035872623324394226 |

Although minutely less than the training and validation results, the test results of 5.4 still surpassed 5.3, with an improvement of around 0.01 for loss and 0.4% for accuracy. While these improvements are miniscule, they are improvements nevertheless, providing confirmation that slightly reducing the augmentation is a step in the right direction.

|  |  |
| --- | --- |
| Test Loss | 0.06092633306980133 |
| Test Accuracy | 0.9854227304458618 |

The accuracy and loss graphs provide a visual representation of the outstanding performance of the model during the training and validation stages of 5.4:

A comparison of graphs with text

AI-generated content may be incorrect.

The confusion matrix provides another visual representation of the brilliance of the model produced during 5.4, as it only mislabelled 9 images. This is 5 images less than 5.3, and another high score in terms of true positive percentage.

A chart with different colored squares

AI-generated content may be incorrect.

Additionally, the minute increase in performance is reflected in the classification report. Although it appears identical to its predecessor, the averages for precision, recall and F1-score are marginally higher than those of 5.3.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class | Precision | Recall | F1-Score | Support |
| Bluebell | 0.9666666666666667 | 0.9666666666666667 | 0.9666666666666667 | 30.0 |
| Buttercup | 1.0 | 1.0 | 1.0 | 38.0 |
| Coltsfoot | 1.0 | 1.0 | 1.0 | 45.0 |
| Cowslip | 0.8823529411764706 | 1.0 | 0.9375 | 30.0 |
| Crocus | 0.9333333333333333 | 1.0 | 0.9655172413793104 | 42.0 |
| Daffodil | 0.96875 | 0.96875 | 0.96875 | 32.0 |
| Daisy | 1.0 | 1.0 | 1.0 | 44.0 |
| Dandelion | 1.0 | 1.0 | 1.0 | 50.0 |
| Fritillary | 1.0 | 1.0 | 1.0 | 48.0 |
| Iris | 1.0 | 1.0 | 1.0 | 35.0 |
| Lilyvalley | 1.0 | 0.9736842105263158 | 0.9866666666666667 | 38.0 |
| Pansy | 1.0 | 0.9302325581395349 | 0.963855421686747 | 43.0 |
| Snowdrop | 1.0 | 1.0 | 1.0 | 37.0 |
| Sunflower | 1.0 | 1.0 | 1.0 | 45.0 |
| Tigerlily | 1.0 | 0.9777777777777777 | 0.9887640449438202 | 45.0 |
| Tulip | 1.0 | 0.9487179487179487 | 0.9736842105263158 | 39.0 |
| Windflower | 1.0 | 1.0 | 1.0 | 45.0 |
| macro avg | 0.9853589965397923 | 0.9862252448134261 | 0.9853767206982075 | 686.0 |
| weighted avg | 0.9878580003429944 | 0.9868804664723032 | 0.9870027796454705 | 686.0 |

By implementing less intense augmentation in 5.4, improvements were made across all metrics, implying that the adjustment made was a success. Performance data as edged even closer to perfection, however there is still a little room for improvement.

5.5

Following the success of increasing the volume of the input dataset from 1x to 5x, stage 5.5 will take this exploration further, by increasing the factor of multiplication to 10. Since the results of 5.4 were marginally better than 5.3, and the lesser degree of augmentation was superior, the degree of augmentation in 5.5 will be the same as its predecessor, the only difference being the aforementioned increased multiplication factor. It is predicted that by providing more high-quality data for training, the model produced will achieve even better results.

mobilenet\_LR1e-05\_BS32\_E30 – HYPERPARAMETER CHAT

In order to compensate for increasing the input dataset, the learning rate has been reduced to 0.00005, and the number of epochs has been increased to 30. This will hopefully allow for a more gradual, thorough and consistent learning process.

After analysing the training results below, it appears that the prediction made above was correct: by providing twice the volume in comparison to the previous stage, the model was able to achieve better results. In the 2nd epoch, a training accuracy of 92.0% was achieved, while from the 4th epoch onwards, the validation accuracy remained north of 99%. The loss data is equally as impressive: from the 6th epoch onwards, the training loss remains below 0.1, while from the 13th epoch onwards, the validation loss was less than 0.02. Remarkably, in the final 6 epochs, the validation loss actually remained below 0.01. Once again, with a new stage, new high scores have been achieved.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Epoch | accuracy | loss | val\_accuracy | val\_loss |
| 1 | 0.720223069190979 | 0.9252597689628601 | 0.949999988079071 | 0.24670648574829102 |
| 2 | 0.9189814925193787 | 0.28397542238235474 | 0.9786764979362488 | 0.12890176475048065 |
| 3 | 0.9481270909309387 | 0.19273102283477783 | 0.9856617450714111 | 0.08291604369878769 |
| 4 | 0.9622264504432678 | 0.13700279593467712 | 0.9926470518112183 | 0.059462714940309525 |
| 5 | 0.9680134654045105 | 0.11401835083961487 | 0.9937499761581421 | 0.048569709062576294 |
| 6 | 0.9753788113594055 | 0.09390944242477417 | 0.9955882430076599 | 0.039764150977134705 |
| 7 | 0.9769570827484131 | 0.08915504068136215 | 0.9955882430076599 | 0.03469324856996536 |
| 8 | 0.9782196879386902 | 0.07676397264003754 | 0.9959558844566345 | 0.02802014909684658 |
| 9 | 0.9822180271148682 | 0.06924855709075928 | 0.9963235259056091 | 0.025619052350521088 |
| 10 | 0.9796927571296692 | 0.06791752576828003 | 0.9970588088035583 | 0.022172439843416214 |
| 11 | 0.9819023609161377 | 0.06351611018180847 | 0.9970588088035583 | 0.022720884531736374 |
| 12 | 0.9829545617103577 | 0.06131436303257942 | 0.9959558844566345 | 0.02010742388665676 |
| 13 | 0.9805344939231873 | 0.0621052160859108 | 0.9977940917015076 | 0.019726401194930077 |
| 14 | 0.9827440977096558 | 0.0563134029507637 | 0.997426450252533 | 0.01621229015290737 |
| 15 | 0.9826388955116272 | 0.05621267110109329 | 0.997426450252533 | 0.014518540352582932 |
| 16 | 0.9836910963058472 | 0.05271307751536369 | 0.9977940917015076 | 0.012878702953457832 |
| 17 | 0.9817971587181091 | 0.05311378464102745 | 0.997426450252533 | 0.014269452542066574 |
| 18 | 0.9860059022903442 | 0.04794914647936821 | 0.9963235259056091 | 0.016370059922337532 |
| 19 | 0.9862163066864014 | 0.04570940509438515 | 0.9985294342041016 | 0.010918369516730309 |
| 20 | 0.9859007000923157 | 0.04356583580374718 | 0.998161792755127 | 0.012143265455961227 |
| 21 | 0.9865319728851318 | 0.04298747330904007 | 0.9988970756530762 | 0.008486246690154076 |
| 22 | 0.9866372346878052 | 0.04276638478040695 | 0.997426450252533 | 0.011019202880561352 |
| 23 | 0.9843223690986633 | 0.04735783115029335 | 0.9985294342041016 | 0.011176807805895805 |
| 24 | 0.9851641654968262 | 0.04437693953514099 | 0.998161792755127 | 0.009232908487319946 |
| 25 | 0.9859007000923157 | 0.043020255863666534 | 0.9988970756530762 | 0.007596189621835947 |
| 26 | 0.9872685074806213 | 0.04164636507630348 | 0.9988970756530762 | 0.00683439988642931 |
| 27 | 0.9893729090690613 | 0.034871093928813934 | 0.9992647171020508 | 0.00801680888980627 |
| 28 | 0.9882155060768127 | 0.03970054164528847 | 0.9992647171020508 | 0.0062927426770329475 |
| 29 | 0.9850589036941528 | 0.04143829643726349 | 0.9985294342041016 | 0.008067913353443146 |
| 30 | 0.9860059022903442 | 0.04153968393802643 | 0.9992647171020508 | 0.00635097362101078 |

New high scores were also obtained during testing, as can be seen in the table below: 0.01 is the lowest loss score ever achieved, and the accuracy of 99.85% means that room for improvement has become even smaller.

|  |  |
| --- | --- |
| Test Loss | 0.010740709491074085 |
| Test Accuracy | 0.9985401630401611 |

Although the training and validation plots of the accuracy and loss graphs appear to have a slight gap between them, this is not being classed as overfitting, as statistically the distance between the plots is miniscule, and is exaggerated by the scale of the graphs. The visible plateaus are also a welcome sight, as they confirm that the model has been trained thoroughly, while its performance remains consistent.

One note worth making from both the table above and the graph below is that peak values are reached within the first 10 epochs. While it is good that they remained at their peaks for the remainder of the training period, 30 epochs may be excessively high, as no benefit is gained during the latter epochs.

A comparison of a graph

AI-generated content may be incorrect.

Impact of the increased dataset is brilliantly reflected in the confusion matrix. Even though the input data was increased 10x compared to the original, and 5x compared to the previous stage, this confusion matrix contains only 2 incorrect predictions. Clearly, the additional input data allowed the model to gain even deeper understanding of the differentiating details of each class, enabling it to achieve near-perfect results.

A chart with different colored squares

AI-generated content may be incorrect.

Once again, even though the difference is minute, the classification report for 5.5 reflects yet another improvement in model performance. 1.0 is the most populous value within the table, reiterating the almost perfect analysis of the confusion matrix, while the remaining values such as the averages at the bottom of the table edge even closer to total accuracy.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class | Precision | Recall | F1-Score | Support |
| Bluebell | 0.9863013698630136 | 1.0 | 0.993103448275862 | 72.0 |
| Buttercup | 1.0 | 1.0 | 1.0 | 75.0 |
| Coltsfoot | 1.0 | 1.0 | 1.0 | 75.0 |
| Cowslip | 1.0 | 1.0 | 1.0 | 81.0 |
| Crocus | 1.0 | 0.9859154929577465 | 0.9929078014184397 | 71.0 |
| Daffodil | 0.9879518072289156 | 1.0 | 0.9939393939393939 | 82.0 |
| Daisy | 1.0 | 1.0 | 1.0 | 95.0 |
| Dandelion | 1.0 | 1.0 | 1.0 | 90.0 |
| Fritillary | 1.0 | 1.0 | 1.0 | 90.0 |
| Iris | 1.0 | 1.0 | 1.0 | 89.0 |
| Lilyvalley | 1.0 | 1.0 | 1.0 | 69.0 |
| Pansy | 1.0 | 1.0 | 1.0 | 69.0 |
| Snowdrop | 1.0 | 1.0 | 1.0 | 78.0 |
| Sunflower | 1.0 | 1.0 | 1.0 | 82.0 |
| Tigerlily | 1.0 | 1.0 | 1.0 | 92.0 |
| Tulip | 1.0 | 0.987012987012987 | 0.9934640522875817 | 77.0 |
| Windflower | 1.0 | 1.0 | 1.0 | 83.0 |
| macro avg | 0.9984854810054077 | 0.9984075576453373 | 0.9984361585836046 | 1370.0 |
| weighted avg | 0.9985589392867943 | 0.9985401459854014 | 0.9985399010990842 | 1370.0 |

5.6

Although the results of 5.5 were close to perfection, there is still a small amount of room for improvement. Stage 5.6 will hopefully conclude the model development process, as the final point of interest is explored, that being the fine-tuning of the base model. Throughout all previous stages of phase 5, the pre-trained *MobileNetV2* architecture had been frozen, meaning that the values within its layers were not manipulated, it was only the additional layers of the network whose values could be adjusted to improve performance. 5.6 is different, as the final 20 layers of the *MobileNetV2*’s architecture have been unfrozen, allowing them to be fine-tuned and optimized to produce a model with even better performance overall.

Due to the excessively long training process in 5.5, the number of epochs has been reduced to 15 for 5.6. To balance this and ensure efficient convergence, the learning rate has been increased to 0.00005.

mobilenet\_finetune\_LR5e-05\_BS32\_E15

The table below contains the performance results of the training process of 5.6. It appears that implementing the fine-tuning was a success, as from around the 3rd and 4th epochs onwards, both the training and validation accuracy is effectively 100%. Additionally, from around a similar point onwards, the training and validation loss values are negligibly minute, with the final training accuracy being 0.002 and the final validation accuracy being 0.0009. This training process has clearly produced the best results of any stage or test throughout the entire project.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Epoch | accuracy | loss | val\_accuracy | val\_loss |
| 1 | 0.7287458181381226 | 0.9673189520835876 | 0.8632352948188782 | 0.4471583068370819 |
| 2 | 0.9644360542297363 | 0.15753474831581116 | 0.96875 | 0.11574357002973557 |
| 3 | 0.9888467788696289 | 0.06195348873734474 | 0.9915441274642944 | 0.04180223494768143 |
| 4 | 0.9947390556335449 | 0.032655052840709686 | 0.998161792755127 | 0.016980042681097984 |
| 5 | 0.997053861618042 | 0.020272744819521904 | 0.9996323585510254 | 0.007962165400385857 |
| 6 | 0.9988425970077515 | 0.013228721916675568 | 1.0 | 0.004651791416108608 |
| 7 | 0.9987373948097229 | 0.010243386961519718 | 0.9996323585510254 | 0.003601826261729002 |
| 8 | 0.9990530014038086 | 0.007150378543883562 | 0.9996323585510254 | 0.0026229710783809423 |
| 9 | 0.9993686676025391 | 0.005513632670044899 | 1.0 | 0.001840853481553495 |
| 10 | 1.0 | 0.0035256489645689726 | 0.9996323585510254 | 0.0019930799026042223 |
| 11 | 0.9992634654045105 | 0.004370170179754496 | 1.0 | 0.001312418025918305 |
| 12 | 0.9994739294052124 | 0.0033531379885971546 | 1.0 | 0.0010911185527220368 |
| 13 | 1.0 | 0.0021787055302411318 | 0.9996323585510254 | 0.0010156809585168958 |
| 14 | 0.9998947978019714 | 0.0016533697489649057 | 0.9996323585510254 | 0.0009210658608935773 |
| 15 | 0.9996843338012695 | 0.0023378911428153515 | 0.9996323585510254 | 0.0009553879499435425 |

The outstanding training process is reflected in equally outstanding test results, as the test accuracy is 100%, while the test loss is 0.0003.

|  |  |
| --- | --- |
| Test Loss | 0.000334896583808586 |
| Test Accuracy | 1.0 |

The graphs accompanying the training process display the perfect outcome of the training process, with rapid convergence and negligible fluctuation.

A graph of a graph of a graph

AI-generated content may be incorrect.

Further confirmation of the success of this training stage comes in the form of the perfect confusion matrix below, which contains not a single incorrect prediction.

A chart with different colored squares

AI-generated content may be incorrect.

As expected, the accompanying classification report echoes the rest of the performance metrics, with the precision, recall and F1-Score of every single class achieving a score of 1.0, or perfection.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class | Precision | Recall | F1-Score | Support |
| Bluebell | 1.0 | 1.0 | 1.0 | 72.0 |
| Buttercup | 1.0 | 1.0 | 1.0 | 75.0 |
| Coltsfoot | 1.0 | 1.0 | 1.0 | 76.0 |
| Cowslip | 1.0 | 1.0 | 1.0 | 84.0 |
| Crocus | 1.0 | 1.0 | 1.0 | 74.0 |
| Daffodil | 1.0 | 1.0 | 1.0 | 80.0 |
| Daisy | 1.0 | 1.0 | 1.0 | 93.0 |
| Dandelion | 1.0 | 1.0 | 1.0 | 88.0 |
| Fritillary | 1.0 | 1.0 | 1.0 | 93.0 |
| Iris | 1.0 | 1.0 | 1.0 | 90.0 |
| Lilyvalley | 1.0 | 1.0 | 1.0 | 74.0 |
| Pansy | 1.0 | 1.0 | 1.0 | 69.0 |
| Snowdrop | 1.0 | 1.0 | 1.0 | 81.0 |
| Sunflower | 1.0 | 1.0 | 1.0 | 80.0 |
| Tigerlily | 1.0 | 1.0 | 1.0 | 93.0 |
| Tulip | 1.0 | 1.0 | 1.0 | 69.0 |
| Windflower | 1.0 | 1.0 | 1.0 | 79.0 |
| macro avg | 1.0 | 1.0 | 1.0 | 1370.0 |
| weighted avg | 1.0 | 1.0 | 1.0 | 1370.0 |

In conclusion, based upon all the data gathered during the training and testing stages, the model created in 5.6 has mastered the input dataset, and has become highly confident and accurate in predicting the 17 species of flowers contained within. Of course, this perfect accuracy is unlikely to remain intact once the model is given new images from external sources such as user photographs, however it has clearly understood the nuances which differentiate the various classes of the dataset, and should retain a solid capability to classify the species it knows with confidence.

Now that perfect results have been achieved by the model of 5.6, the model development stage can be relievingly concluded. The next stage of the project will be to create some kind of application to house the model, so that users will be able to interact with it, providing it with images of flowers and receiving predictions on their species. Once the app is created, both the app and model can be tested in real-world scenarios.