SUPPORTING INFORMATION
Unsupervised Structure Classes *vs* Supervised Property Classes of Silicon Quantum Dots using Neural Networks

Amanda J. Parker[1] and Amanda S. Barnard*[2]

[1]*Data61 CSIRO, Docklands 3008, Australia*
[2]*Research School of Computer Science, Australian National University, Acton 2601, Australia*

The document contains further information and results to support the conclusions in the main text, including the correlations matrix and skew map in Fig. S1. The distribution of the supervised property classes (color bands) with respect to each of the structural features is shown in Fig. S2, and the distribution of the unsupervised property classes (ILS clusters) with respect to each of the structural features is shown in Fig. S3. The weights and biases for the neural network classifier for the unsupervised structure classes are provided, along with the neural network regressor, which can be used to construct an analytical expression for classifying surface passivated SiQDs based on the the size and shape. The weights and biases for the neural network classifier for the supervised property classes are also provided, along with results using a random forest classifier in Fig. S4. A flow chart of the entire workflow is shown in Fig. S5.

*Corresponding author: amanda.s.barnard@anu.edu.au

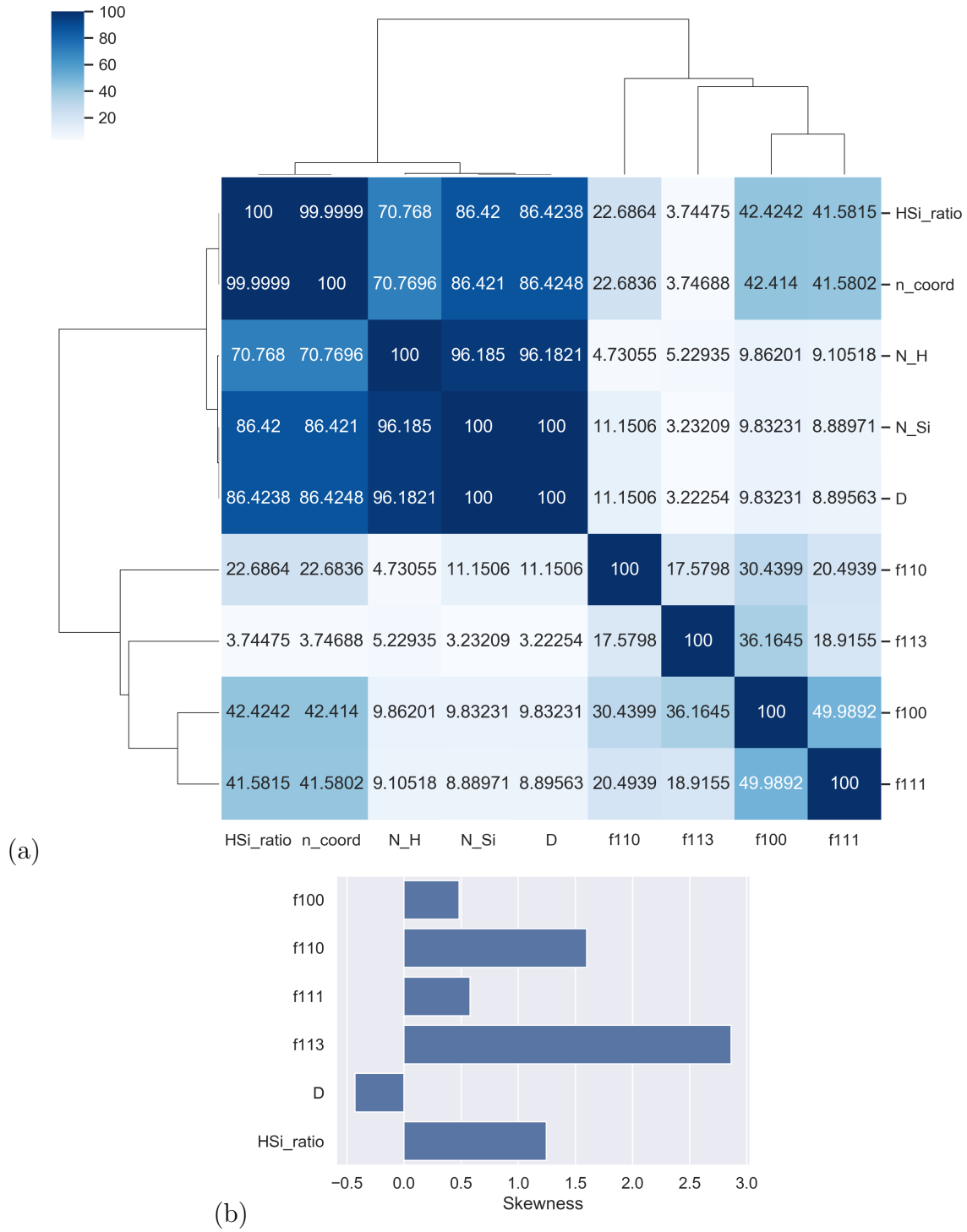

Figure S1: (a) The correlation matrix for the entire SiQD data set, and (b) the skew maps of the retained features.

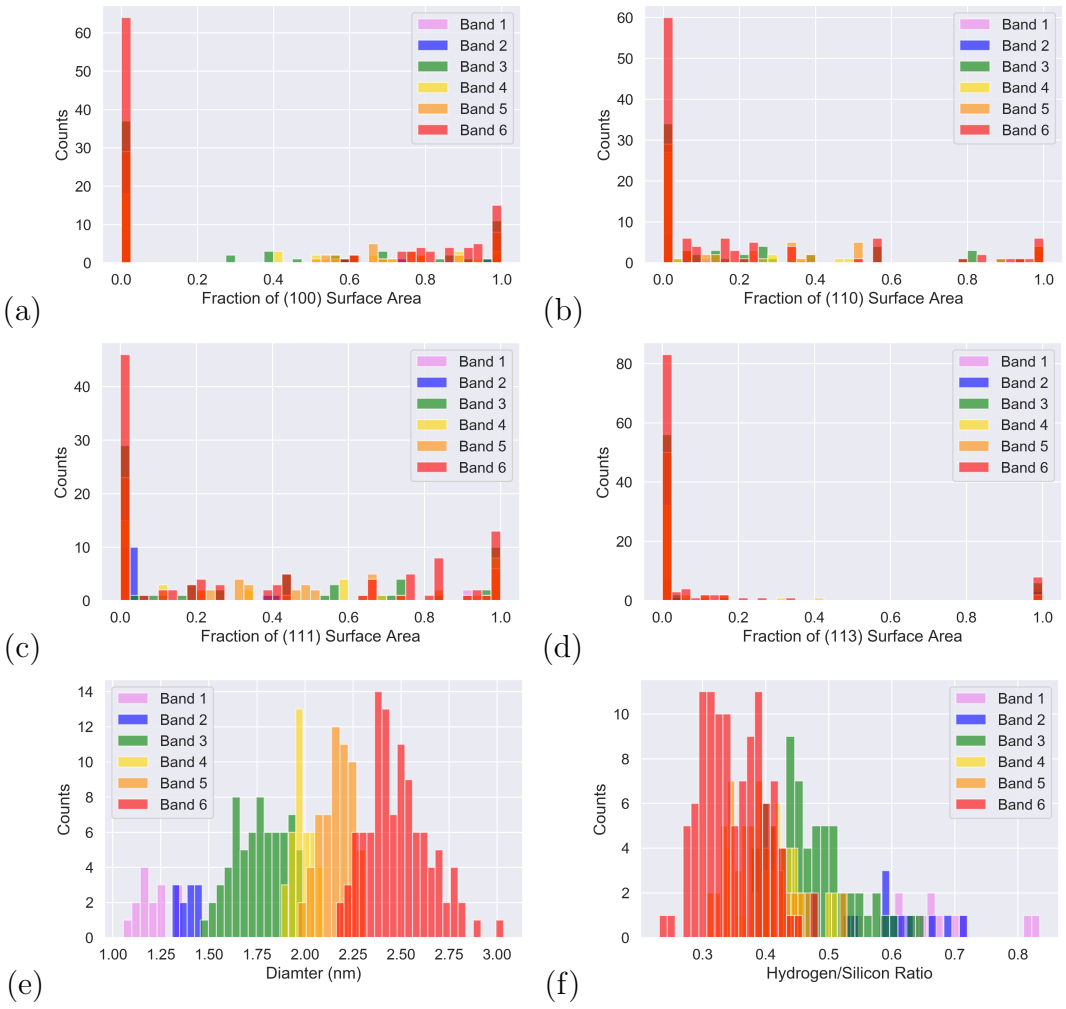

Figure S2: Distribution of the supervised property classes with respect to (a) the fraction of {100} surface area, (b) the fraction of {110} surface area, (c) the fraction of {111} surface area, (d) the fraction of {113} surface area, (e) the average particle diameter (in nm) and , (f) the hydrogen to silicon ratio.

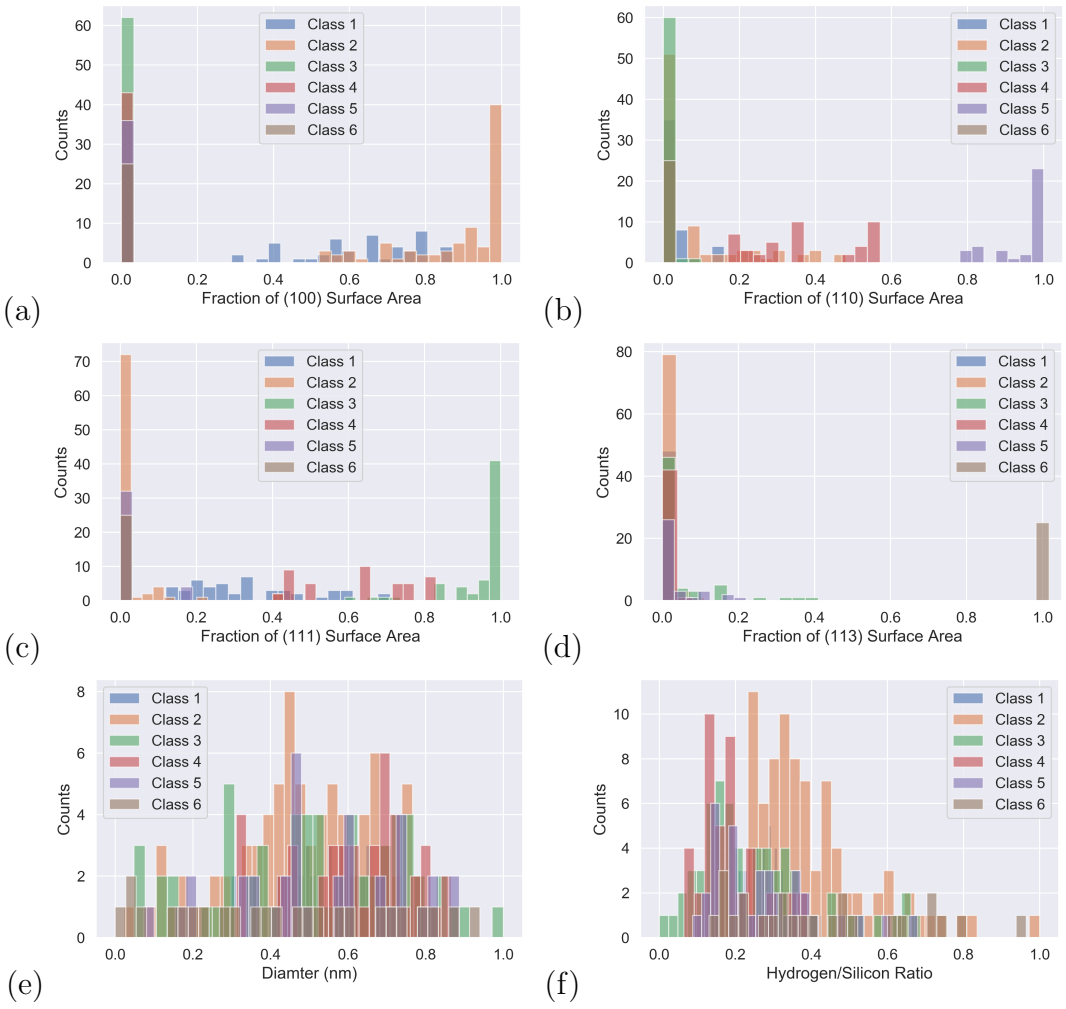

Figure S3: Distribution of the unsupervised structure classes with respect to (a) the fraction of {100} surface area, (b) the fraction of {110} surface area, (c) the fraction of {111} surface area, (d) the fraction of {113} surface area, (e) the average particle diameter (in nm), and (f) the hydrogen to silicon ratio.

Neural Network Weights and Biases

Unsupervised Structure Classes

The weights and biases for the one hidden layer and the output layer for unsupervised structure classes, and the output layer for each class. Format: hidden layer, neuron [w_i ... w_I],Bias

```
Hidden 0, 0 [-0.982 -2.759  1.334  2.218 -0.149 -2.265],+8.776
Hidden 0, 1 [ 4.201 -2.75  -1.066  0.061 -1.533  1.062],+4.215
Hidden 0, 2 [ 0.016 -0.031 -0.034 -0.022 -0.005 -0.015],-0.465
Hidden 0, 3 [-0.383  2.302 -0.016 -1.901  0.394  1.302],+12.329

Hidden 1, 0 [1.844 0.12  0.039 0.343],-0.695
Hidden 1, 1 [-2.811 -4.003  0.026  3.077],+10.043
Hidden 1, 2 [-3.098  3.65  -0.016  1.459],-2.526

Output, 0 [ 0.532 -1.994 -0.22 ],+30.405
Output, 1 [-1.132  0.214  4.51 ],-34.424
Output, 2 [ 0.806 -0.486 -1.271],+16.349
Output, 3 [ 0.162  1.969 -1.19 ],-8.65
Output, 4 [-1.005  3.78  -0.598],-43.116
Output, 5 [ 0.653 -3.381 -1.307], +38.491
```

Supervised Property Classes

The weights and biases for the three hidden layers and the output layer for supervised property classes, and the output layer for each class. Format: hidden layer, neuron [w_i ... w_I],Bias

```
Hidden 0, 0 [-0.949  0.385  0.305  0.38  -1.139 -0.968],+8.650
Hidden 0, 1 [-0.016  0.008 -0.004  0.006 -0.416 -0.332],-2.091
Hidden 0, 2 [-0.05   0.121  0.051  0.084 -1.488 -1.122],+7.004
Hidden 0, 3 [-0.009 -0.09  -0.029 -0.088  1.957  1.45 ],-5.279
Hidden 0, 4 [-0.13   0.086  0.046  0.063 -0.733 -0.581],+1.139
Hidden 0, 5 [ 0.627 -0.141 -0.208 -0.188 -1.005 -0.761],-5.025
Hidden 0, 6 [-0.294  0.005  0.065  0.038  1.709  1.258],-2.874
Hidden 0, 7 [ 0.858 -0.269 -0.264 -0.279 -0.702 -0.451],-5.937
Hidden 0, 8 [-0.132  0.048  0.009  0.058 -0.842 -0.674],-4.360
Hidden 0, 9 [-0.462  0.169  0.172  0.169 -0.339 -0.259],+6.246

Hidden 1, 0 [ 0.422  0.009 -0.049  0.104  0.048 -0.33   0.225 -0.453  0.044  0.201],+3.246
Hidden 1, 1 [-0.651 -0.201 -0.495  0.666 -0.297 -0.206  0.485 -0.037 -0.433 -0.219],-3.856
```

```
Hidden 1, 2 [-1.258 -0.358 -1.373  1.755 -0.686 -0.763  1.48  -0.412 -0.742 -0.412],-2.996
Hidden 1, 3 [-0.802 -0.272 -1.023  1.312 -0.501 -0.67   1.148 -0.437 -0.546 -0.239],-1.729
Hidden 1, 4 [ 0.335 -0.008 -0.279  0.342 -0.022 -0.449  0.458 -0.55   0.06   0.151],+4.802
Hidden 1, 5 [-0.598 -0.079 -0.306  0.366 -0.185  0.077  0.199  0.219 -0.193 -0.251],-2.077
Hidden 1, 6 [-0.132 -0.117 -0.643  0.863 -0.251 -0.617  0.862 -0.602 -0.219 -0.008],+2.924

Hidden 2, 0 [-0.405 -0.507 -1.435 -1.167 -0.575 -0.063 -0.976],-4.655
Hidden 2, 1 [0.336 0.763 2.303 1.799 0.701 0.304 1.312],-3.782
Hidden 2, 2 [ 0.562 -0.903 -1.878 -1.214  0.371 -0.856 -0.296],+1.156

Output, Band 1 [ 1.778 -1.994  0.801],-6.929
Output, Band 2 [ 0.522 -1.248  0.82 ],1.258
Output, Band 3 [-0.222 -0.604  0.676],+6.703
Output, Band 4 [-0.279  0.591  0.315],+3.092
Output, Band 5 [-0.554  1.57  -0.401],-3.268
Output, Band 6 [-1.239  1.687 -2.217],+1.832
```

Supervised Regression

The weights and biases for the three hidden layers and the output layer for supervised property classes, and the output layer for each class. Format: hidden layer, neuron [w_i ... w_I],Bias

```
Hidden 0, 0 [-0.107 -0.41   0.096 -0.176 -0.606  0.184],-0.422
Hidden 0, 1 [ 0.449  0.469 -0.1   -0.208  1.031  1.08 ],+0.324
Hidden 0, 2 [ 0.864  0.774  0.08  -0.458 -1.551  1.773],+1.916
Hidden 0, 3 [0.457 0.249 0.186 0.453 2.933 3.991],+1.967
Hidden 0, 4 [ 0.712 -1.385 -0.179  1.198 -1.212  1.085],+0.440
Hidden 0, 5 [-0.128 -0.041  0.313  0.407  1.489  2.53 ],+1.147
Hidden 0, 6 [-0.378 -0.167 -0.257 -0.167 -0.161 -0.352],-0.440
Hidden 0, 7 [ 0.669  0.454  0.465 -0.642  1.8    1.325],+1.860
Hidden 0, 8 [ 0.086 -0.058  0.079  0.158  0.04  -0.389],-0.521
Hidden 0, 9 [ 0.035  0.043  0.792  0.683 -2.126  1.376],+3.030

Hidden 1, 0 [ 0.258  0.299  0.257 -0.365 -0.16  -0.166  0.367  0.202  0.257 -0.069],-0.637
Hidden 1, 1 [-0.269  0.889 -2.903  4.75  -2.289  2.876 -0.026  1.531 -0.006 -3.28 ],+5.021
Hidden 1, 2 [-0.469 -0.135 -0.15  -2.    -0.18  -1.085 -0.361 -0.793  0.022 -0.381],-3.131

Output, 0 [ 0.015  7.314 -0.022],+4.578
```

Class	Precision	Recall	F1-score	SiQDs
Band 1	0.50	0.67	0.57	15
Band 2	0.00	0.00	0.00	11
Band 3	0.60	0.82	0.69	67
Band 4	0.50	0.33	0.40	38
Band 5	0.90	1.00	0.95	58
Band 6	1.00	1.00	1.00	108

Table S1: Classification reports for the supervised property classes separated using the optimized Random Forest classifier.

Supervised Property Classes with the Random Forest Classification

The supervised property classes where separated using Random Forest classifier with the optimized hyper-parameters: {criterion = 'entropy', max_depth=13, min_samples_leaf=6, min_samples_split=10, min_weight_fraction_leaf = 0.0336, max_features=None, n_estimators=493, random_state=42}. Using these hyper-parameters the training score for the supervised classes was $R^2_{train} = 0.924$, the testing score was $R^2_{test} = 0.817$ and the cross validation score based on 10-fold cross validation was $R^2_{CV} = 0.8 \pm 0.148$.

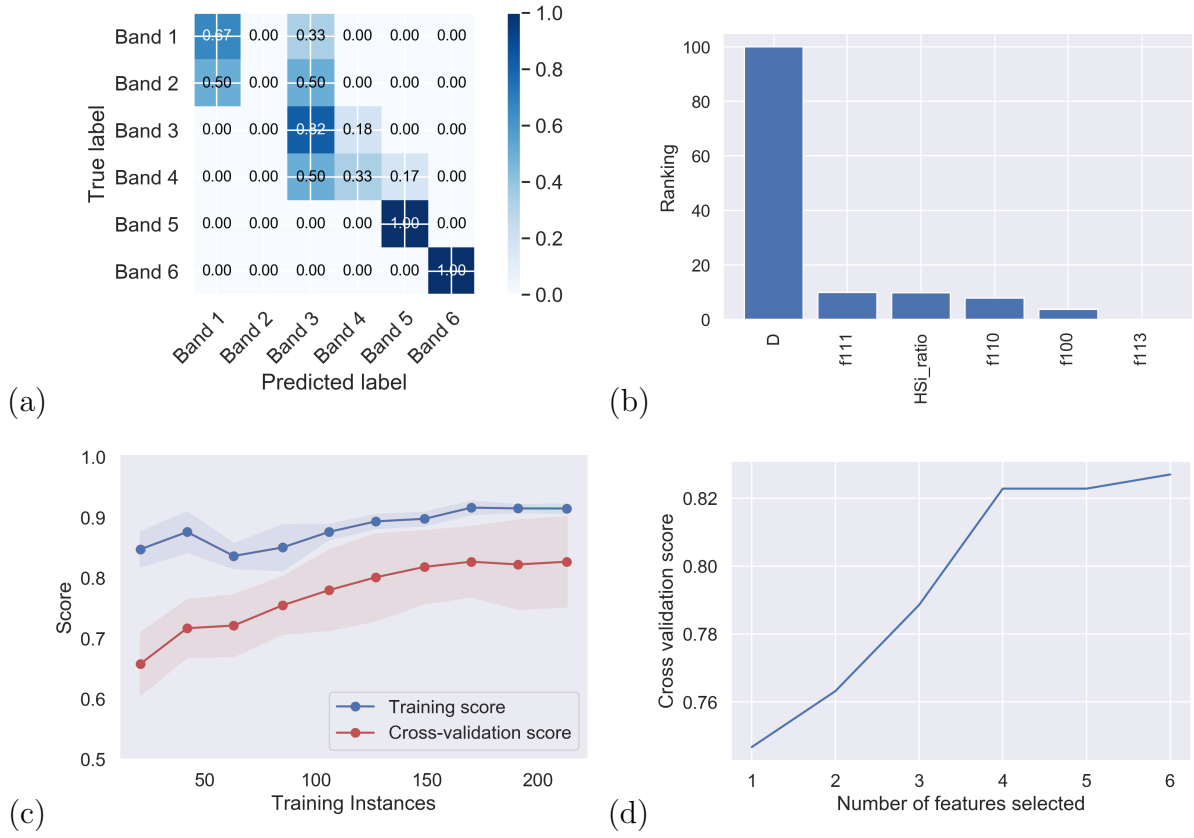

Figure S4: (a) Confusion matrix for the supervised property classes separated using the Random Forest classifier, (b) the feature importance profile, (c) the learning curve showing the overall accuracy and generalizability, and (d) the recursive feature elimination plot showing even the (unimportant) {113} facets are necessary to the model.

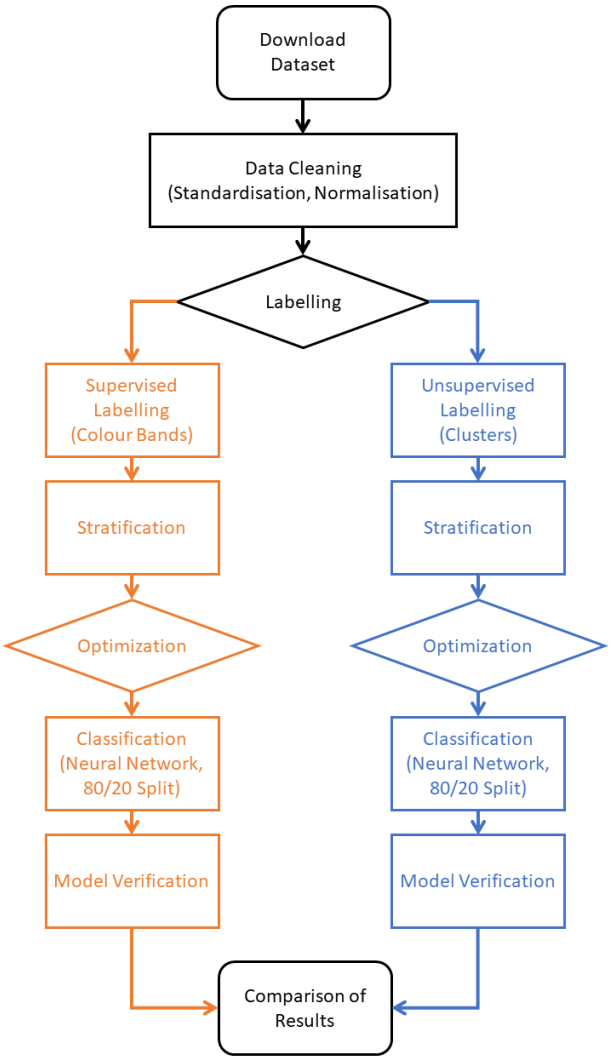

Figure S5: A flow chart of the workflow used in this study. Machine learning and data science packages were used from scikit-learn.