**Supervised machine learning classifies inflammatory bowel disease patients by subtype using whole exome sequencing data**

**SUPPLEMENTARY DATA**

Hyperparameter Tuning Value Options

Minimum samples per leaf = 1, 2, 3, 4, 5, 6

Maximum Features = 'sqrt', 'log2', None

Minimum samples per split = 2, 3, 4, 5, 6

Maximum depth = 1-30, None

Number of estimators = 100, 250, 500, 750, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000

Hyperparameter Tuning Supplementary Tables

Supplementary Table 1: Nested Cross Validation of Hyperparameters on **all available genes**, 7 fold outer CV, 5 fold inner CV. Hyperparameter values chosen from fold 3 for final classifier.

|  |  |  |  |
| --- | --- | --- | --- |
| **Fold** | **Outer Fold Balanced Accuracy** | **Inner CV Balanced Accuracy** | **Optimal Hyperparameters** |
| **Max Depth** | **Max Features** | **Minimum Samples per Leaf** | **Minimum Samples per Split** | **Number of Estimators** |
| 1 | 0.85985 | 0.78155 | 22 | Log2 | 1 | 2 | 7000 |
| 2 | 0.726916 | 0.729369 | 29 | sqrt | 2 | 4 | 3000 |
| 3 | 0.755952 | 0.753946 | 29 | sqrt | 1 | 4 | 10,000 |
| 4 | 0.834704 | 0.774565 | 25 | Log2 | 1 | 4 | 9000 |
| 5 | 0.919768 | 0.762910 | 23 | Log2 | 1 | 3 | 9000 |
| 6 | 0.680944 | 0.753255 | 23 | Log2 | 1 | 5 | 6000 |
| 7 | 0.788254756 | 0.756290 | 15 | Log2 | 1 | 6 | 10,000 |

Supplementary Table 2: Nested Cross Validation of Hyperparameters on **autoimmune gene panel**, 7 fold outer CV, 5 fold inner CV. Hyperparameter values chosen from fold 4 for final classifier.

|  |  |  |  |
| --- | --- | --- | --- |
| **Fold** | **Outer Fold Balanced Accuracy** | **Inner CV Balanced Accuracy** | **Optimal Hyperparameters** |
| **Max Depth** | **Max Features** | **Minimum Samples per Leaf** | **Minimum Samples per Split** | **Number of Estimators** |
| 1 | 0.55682 | 0.60444 | 5 | None | 2 | 5 | 10000 |
| 2 | 0.56533 | 0.61313 | 10 | None | 4 | 2 | 3000 |
| 3 | 0.500 | 0.6259628013 | 8 | None | 6 | 5 | 100 |
| 4 | 0.5723684 | 0.58872284 | 2 | Sqrt | 3 | 4 | 250 |
| 5 | 0.5632754 | 0.585826257 | 10 | Log2 | 1 | 3 | 7000 |
| 6 | 0.55856643 | 0.585098191748 | 20 | None | 2 | 4 | 500 |
| 7 | 0.6261373035566584 | 0.5713827454283834 | 1 | None | 2 | 3 | 100 |

Supplementary Table 3: Nested Cross Validation of Hyperparameters on **IBD gene panel**, 7 fold outer CV, 5 fold inner CV. Hyperparameter values chosen from fold 5 for final classifier.

|  |  |  |  |
| --- | --- | --- | --- |
| **Fold** | **Outer Fold Balanced Accuracy** | **Inner CV Balanced Accuracy** | **Optimal Hyperparameters** |
| **Max Depth** | **Max Features** | **Minimum Samples per Leaf** | **Minimum Samples per Split** | **Number of Estimators** |
| 1 | 0.501736111 | 0.614874590 | 7 | None | 6 | 2 | 3000 |
| 2 | 0.61019163 | 0.577843950 | 1 | None | 3 | 4 | 10000 |
| 3 | 0.5492160278 | 0.6057919209 | 4 | None | 6 | 4 | 5000 |
| 4 | 0.6026182432 | 0.5974296239 | 3 | None | 3 | 2 | 6000 |
| 5 | 0.62564102564 | 0.61776382549 | 2 | None | 6 | 2 | 4000 |
| 6 | 0.5974955277 | 0.591813603 | 1 | None | 3 | 4 | 750 |
| 7 | 0.4987179487 | 0.6029599539 | 18 | None | 3 | 4 | 100 |



Supplementary Figure 1 Gene GenePy score distributions, and their contributions to the random forest model. A) Distributions of GenePy scores of the top 10 genes in the classifier that utilised all available genes, grouped by IBD subtype. B) SHAP values representing GenePy scores contributions to classification by random forest.



Supplementary Figure 2 Gene GenePy score distributions, and their contributions to the random forest model. A) Distributions of GenePy scores of the top 10 genes in the classifier that utilised the IBD gene panel, grouped by IBD subtype. B) SHAP values representing GenePy scores contributions to classification by random forest.

Supplementary Figure 3- AUROC analysis using *NOD2* only to differentiate between Crohn’s disease and ulcerative colitis. AUC 0.61, with an optimised cut-off of value in this training cohort of 0.2798. Applying this cut-off value to a testing set of data demonstrated a statistically significant (ꭓ2) ability to predict CD vs UC using *NOD2* alone, p=0.003.

****