**Supplementary Material**

All supplementary materials are publicly available and can be accessed on the Cambridge Core website.

**Program script and sample data** All program codes (in R language) of our three ML classifier implementation are made available in our open GitLab repository (*https://gitlab.com/izhardiansyah76/ml-zygosity-classification*). We also provide therein a number of pretrained models which can readily be used either directly for predicting new data or for retraining with additional data, i.e. by adding new rows to our existing sample dataset. The sample dataset we provided here was made completely anonymous to protect our participants’ privacy. Visit the GitLab repository for instructions on how to use the pretrained models, set up your dataset, augment the existing dataset, etc.

**Supplementary Table S1** The four items used in the manual analysis and the scores assigned to their responses.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Item** | **MZ answer** | **Score** | **DZ answer** | **Score** | **Invalid answer** | **Score** |
| Are your twins "like two peas in a pod" or not more similar than siblings in general? | Like two peas in a pod | 1 | Not more alike than siblings in general | -1 | Don’t knowDon’t want to answer | 0 |
| How alike are your twins in terms of eye color? | Exactly alike | 1 | Different | -1 | Somewhat alikeDon’t knowDon’t want to answer | 0 |
| How alike are your twins in terms of hair color? | Exactly alike | 1 | Different | -1 | Somewhat alikeDon’t knowDon’t want to answer | 0 |
| How often do your twins get mixed up by strangers? | Almost always or always | 1 | RarelyAlmost never or never | -1 | OftenDon’t knowDon’t want to answer | 0 |

###

### Supplementary Table S2 Frequency distribution of responses for each of 8 questionnaire items

|  |  |  |
| --- | --- | --- |
| **a. Are the twins like two peas in a pod?** |  | **b. How often do you mix up the twins?** |
| **Choice** | **MZ** | **DZ** |  | **Choice** | **MZ** | **DZ** |
| “Not more alike than siblings in general” | 27 | 125 |  | “Almost never or never” | 0 | 46 |
| “Don’t know” | 17 | 3 |  | “Rarely” | 5 | 53 |
| “Like two peas in a pod” | 110 | 0 |  | “Don’t know” | 3 | 3 |
| “Decline” | 2 | 0 |  | “Often” | 58 | 17 |
|  |  | “Almost always or always” | 90 | 9 |
| **Total** | **156** | **128** |  | **Total** | **156** | **128** |

|  |  |  |
| --- | --- | --- |
| **c. How similar are the twins’ eye colour?** |  | **d. How similar are the twins’ hair colour?** |
| **Choice** | **MZ** | **DZ** |  | **Choice** | **MZ** | **DZ** |
| “Different” | 0 | 41 |  | “Different” | 2 | 45 |
| “Don’t know” | 1 | 1 |  | “Don’t know” | 7 | 5 |
| “Somewhat alike” | 20 | 63 |  | “Somewhat alike” | 22 | 67 |
| “Exactly alike” | 135 | 23 |  | “Exactly alike” | 125 | 11 |
| **Total** | **156** | **128** |  | **Total** | **156** | **128** |

|  |  |  |
| --- | --- | --- |
| **e. Are the twins one-egg or two-egg?** |  | **f. How similar are the twins’ earlobes?** |
| **Choice** | **MZ** | **DZ** |  | **Choice** | **MZ** | **DZ** |
| “Surely two-egg twins” | 5 | 39 |  | “Clearly different” | 2 | 33 |
| “Probably two-egg twins” | 23 | 83 |  | “Somewhat different” | 42 | 49 |
| “Don’t know” | 12 | 6 |  | “Don’t know” | 22 | 23 |
| “Probably one-egg twins” | 63 | 0 |  | “Not different at all” | 90 | 23 |
| “Surely one-egg twins” | 53 | 0 |  |  |
| **Total** | **156** | **128** |  | **Total** | **156** | **128** |

|  |  |  |
| --- | --- | --- |
| **g. Do the twins have the same blood group?** |  | **h. Do the twins have the same rhesus factor?** |
| **Choice** | **MZ** | **DZ** |  | **Choice** | **MZ** | **DZ** |
| “Different” | 0 | 4 |  | “Different” | 0 | 2 |
| “Don’t know” | 133 | 116 |  | “Don’t know” | 137 | 120 |
| “Same” | 23 | 8 |  | “Same” | 19 | 6 |
| **Total** | **156** | **128** |  | **Total** | **156** | **128** |

Response frequencies are cross-tabulated by zygosity class and answer choices, for each questionnaire items separately; Items a to d are the ones used in manual approach and in the final ML model-fitting.

**Supplementary Figure S1** Relative importance of the questionnaire items used for initially training the Random Forest zygosity classifier.



The random forest algorithm we employed here uses the Gini Index as a measure of purity (where a larger index corresponds to less purity, hence a decrease in Gini Index indicates an increase in purity of the split).