PLOS ONE

RESEARCH ARTICLE

Can you make morphometrics work when you know the right answer? Pick and mix approaches for apple identification

Maria D. Christodoulou[1¤], **Nicholas Hugh Battey**[2], **Alastair Culham**[1]*

1 University of Reading Herbarium, Harborne Building, School of Biological Sciences, University of Reading, Whiteknights, Reading, United Kingdom, **2** School of Biological Sciences, University of Reading, Whiteknights Reading, United Kingdom

¤ Current address: Department of Statistics, University of Oxford, Oxford, United Kingdom
* a.culham@reading.ac.uk

Citation: Christodoulou MD, Battey NH, Culham A (2018) Can you make morphometrics work when you know the right answer? Pick and mix approaches for apple identification. PLoS ONE 13(10): e0205357. https://doi.org/10.1371/journal.pone.0205357

Editor: Suzannah Rutherford, Fred Hutchinson Cancer Research Center, UNITED STATES

Received: March 26, 2018

Accepted: September 24, 2018

Published: October 15, 2018

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by BBSRC: 1132848 (NHB), https://www.bbsrc.ac.uk/. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no completing interests exist.

Abstract

Morphological classification of living things has challenged science for several centuries and has led to a wide range of objective morphometric approaches in data gathering and analysis. In this paper we explore those methods using apple cultivars, a model biological system in which discrete groups are pre-defined but in which there is a high level of overall morphological similarity. The effectiveness of morphometric techniques in discovering the groups is evaluated using statistical learning tools. No one technique proved optimal in classification on every occasion, linear morphometric techniques slightly out-performing geometric (72.6% accuracy on test set versus 66.7%). The combined use of these techniques with post-hoc knowledge of their individual successes with particular cultivars achieves a notably higher classification accuracy (77.8%). From this we conclude that even with predetermined discrete categories, a range of approaches is needed where those categories are intrinsically similar to each other, and we raise the question of whether in studies where potentially continuous natural variation is being categorised the level of match between categories is routinely set too high.

Introduction

With more than 7,000 apple cultivars described [1](some authors estimate 10,000 cultivars [2]), fruit of all shapes, sizes, colours, flavour, and texture exist. This diversity makes identification a challenging task. From hominid stone implement design [3] to the identification of fossil sharks from their teeth [4], the extensive development of morphometric tools in the past few decades [5–7], has resulted in many exciting discoveries across scientific disciplines. In areas such as forensics and palaeontology, morphometrics may be the only tool available to researchers [4,8]. For over 2000 years morphology has remained the primary tool for field classification [9] although the tools used to gather data and analyse them have changed substantially. The classification of objects in general is a natural reaction of humans to the complexity of the world that surrounds them. Humans excel at pattern matching [10], a skill often exploited for

security systems [11,12] and essential to classification. Arguably, this tendency can result in pareidolia, the misclassification of features to fit a preconceived model of limited scope [13]. Nevertheless, advanced pattern matching remains a crucial tool for navigating day to day life [14]. Many of the uses of pattern recognition (e.g. number plate reading [15] in carparks) rely heavily on statistical classification techniques, and have many potential biological applications [16–21]. Here we explore non-destructive sampling for classification of apple cultivars, the identity of which traditionally relies on a small number of acknowledged apple experts, usually working with large collections of named and curated apple trees, who have gained years of practical experience of those apples cultivars yet the correct classification of an apple has immediate economic impact. Government figures show the wholesale price of 'Gala' versus 'Braeburn', for instance, can differ by 20% or more per kilo [22].

Continuous development in collection, recording, and analysis methods has given morphology a very sophisticated toolkit for taxonomists. Many recent taxonomic publications have exploited morphology under the umbrella of integrative taxonomy [23] which relies on the use of multiple data sources for inference [23]. The techniques most commonly combined with morphometrics are molecular [24], but can also include cytometry [23,25], chromosome counts [25] or the chemical composition of secreted compounds [26], all of which involve destructive sampling. The most important aspect of integrative taxonomy is the use of the appropriate data sources for the organisms in question. Combination of morphometrics and molecular markers can prove very successful in the delimitation of closely related taxa, both within botanical [27] and zoological [28–30] research. This success is taxon and technique dependent, as illustrated by the absence of morphometric resolving power in the works by Mamos et al. [31] and Lecocq et al. [26]. Diagnostic characters are often difficult to determine and quantify, and the selection process is challenging. Some examples of this difficulty include selecting the appropriate life stage [32]—contrasting larval stages to adults on *Culex* species—or morphological character—contrasting overall shape to specific landmarks in *Cobitis* populations [33]. Although the majority of these examples focus on shape description and quantification, colour may also be a vital source of morphometric data [24,29].

Talented human experts can take years to master apple cultivar identification. By studying both internal and external morphological characters, apple experts rely on their in-depth knowledge of hundreds of cultivars, contextual awareness, and their understanding of biological variation within those cultivars to classify unknown samples [34–36]. They also commonly analyse their observations in a flexible manner, focusing on some aspects of the morphology more heavily in some cases than in others. To illustrate this, we present the hypothetical case of an expert identifying an apple that is uniformly dark red. In that case the expert would not consider cultivars which are almost exclusively green or yellow in colour, such as 'Granny Smith' and 'Golden Delicious', even if the shape and size of the sample fruit matches those cultivars; the expert would simply ignore the similarities in shape and focus on shape characters for apples that can be dark red in colour.

The fundamental challenge in identification of an individual apple by an expert is much greater than that, for instance, of identification of many bird species which can be done routinely at great distance using binoculars, due to the presence of consistent landmarks of shape, size, colour, etc. Fine-grained recognition algorithms are successful in identifying different species of birds in a variety of environments and from a variety of angles because the object being identified is fundamentally consistent in size, shape, and colour [37]. Similarly, the consistency of size, shape, and colour in flowers of the same taxon leads to routine benchmarking of fine-grained algorithms against floral datasets [37,38]. This has caused these characters to be used extensively in plant classification. Even the very well-studied British flora has only recently gained an identification guide that does not depend on features flowers provide [39],

despite the fact that experienced field botanists have long been able to identify plants in a vegetative state through knowledge and intuition. In the case of apples there is a need to identify individual fruit separated from the parent tree. The identification is at the level of cultivar and not species, and therefore the expected level of difference is small. As such it becomes crucial to standardise the imaging approach of the apples, such that variation detected is that of the fruit and not of its surroundings and the angle at which it is viewed.

Apple variety identification provides an ideal model to test the limits of morphological classification in biology because apple cultivars are usually clones and therefore the variation found is likely to be environmental in cause, and not genetic. By analysing clonal cultivars, we can be confident that there is a single correct answer to any identification. Both the challenge and novelty of this work is to discover whether apple cultivars can be identified accurately and reliably based on visual cues alone, in the absence of taste and smell. The challenge closest to our work is the collection of studies by Corney and colleagues [40–42] on automatic classification tools for *Tilia* leaves. The absence of sufficient landmarks for apple cultivars inspired us to study them from first principles, returning to basic morphometric tools and concepts in order to design a classification protocol.

Here we aim to discover whether the currently available arsenal of morphometric approaches is capable of grouping individual apples into their correct cultivar. We demonstrate that through the use of combined approaches a success rate of 78% can be achieved in this particularly challenging biological identification problem.

Materials and methods

Fruit of twenty-seven apple cultivars were collected at the National Fruit Collection in Brogdale, Kent during the 2013 and 2014 growing seasons. These were collected when considered ready to harvest by the professional pickers, who routinely use appearance and flavour as indicators of ripeness. The list of cultivars sampled is presented in S1 Table. Maximum length and maximum diameter were measured for each fruit using Vernier callipers (Mitutoyo Corporation, Japan). Weight, after removal of pedicel, was measured using precision scales calibrated to 0.01g (Denver Instrument S-402, New York). All measurements were made within 24 hours of harvest.

Each apple was placed against a blue (RGB: 0, 0, 255) background on a Kaiser Phototechnik R1 photographic stand and was photographed using a Nikon D5100 camera with a Nikon AF-S 40mm Micro NIKKOR f/2.8 DX G lens. The blue background was selected because it would interact to the smallest degree with apple skin colour, which is predominantly a combination of red and green pixels. The camera was positioned 0.50 m above the base of the stand, a setting that was not altered during the data collection and allowed capture of the entire outline of even the largest apples in the sample, at the same time retaining sufficient resolution for detailed digitisation. Each fruit was photographed a total of six times (Fig 1): one image for the calyx end, one for the pedicel end and four side-images (fruit rotated by 90˚ clockwise for every image), resulting in a total of 3,240 images (original image dimensions 4928x3264 pixels). Of the four side-images per fruit only the first two (the original and the 90˚ rotation from the original) were unique in terms of shape, the other two being their mirror images.

On each image, landmarks were recorded manually using the tpsDig2 software [43]. Landmark selection relied on the ability to consistently obtain the same landmarks on all the fruit. By observing collections of images from each cultivar, six landmarks were selected for the digitisation: two on the crown apices, two on the shoulder apices, one on the calyx and one on the pedicel attachment point (illustrated in Fig 2).

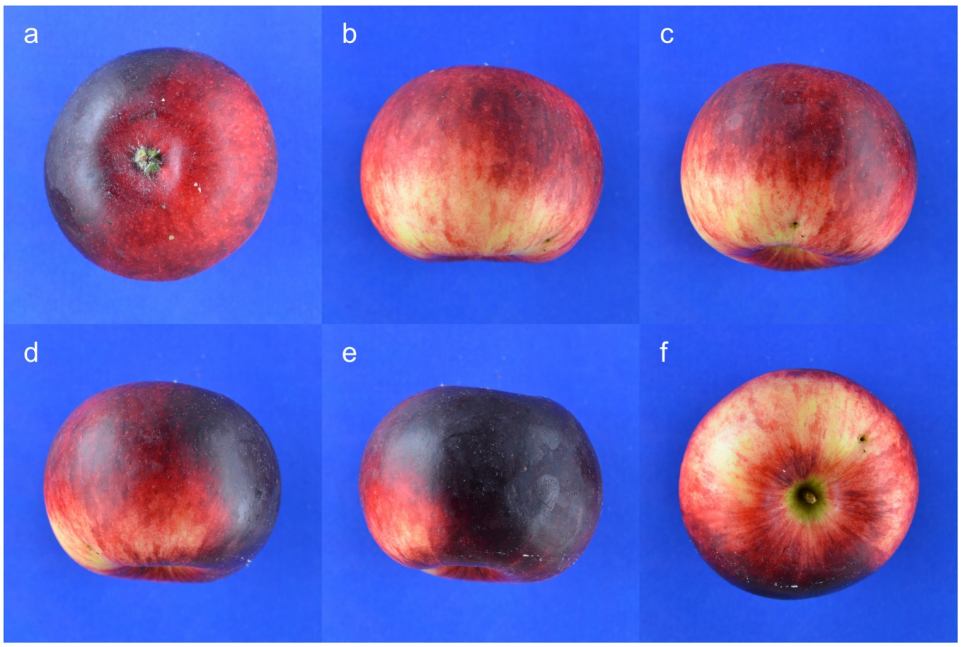

Fig 1. Example of all six captured images for one apple. a) calyx, b-e) side-views, each at 90˚ to each other, f) pedicel.

https://doi.org/10.1371/journal.pone.0205357.g001

To establish the degree of digitisation error, all 3,240 images were digitised twice, with a two-week gap, to ensure that the second digitisation was not affected by muscle memory. Analysis of digitisations was conducted using MorphoJ [44]. Digitisation error was calculated using Procrustes ANOVAs and found to be negligible across all samples. The Procrustes ANOVAs for the two separate digitisations had a smaller and significantly different mean square error estimate for digitisation than for individuals, suggesting that digitisation error was negligible. The result was similar for the comparison of the two sets of side-view images (original and 90˚ versus 180˚ and 270˚) confirming that the digitisation error was negligible.

The first image of each fruit could have been used exclusively to describe its shape. This, however, would ignore the variation that the 90˚ rotation could provide. To be able to include the variation from the two views as well as to standardise between fruit, the landmark positions from the two views after a Procrustes superimposition were averaged. This process was repeated for the 180˚ and 270˚ views and the two datasets were then compared to establish possible digitisation error, which was also found to be negligible. After the Procrustes superimposition, the centroid size for each fruit was recorded. The Procrustes coordinates were then used to perform a Principal Components Analysis (PCA), the scores from which were recorded for each fruit.

Colour measurements were obtained by estimating the overall Red, Green and Blue (RGB) intensities per pixel for each image using ImageJ [45]. To reduce the dimensionality of the RGB colour measurements and remove the variation caused by the auto-white balance, a PCA was performed and the first principal component was retained as the overall colour measurement. The calyx images for each fruit were used to measure calyx area and the calyx "eye" (an opening in the calyx). This was performed using the tpsDig2 [43] by manually outlining the relevant edges.

From these measurements, two datasets were compiled, a linear morphometrics dataset including: maximum length; maximum diameter; weight; first principal component of colour;

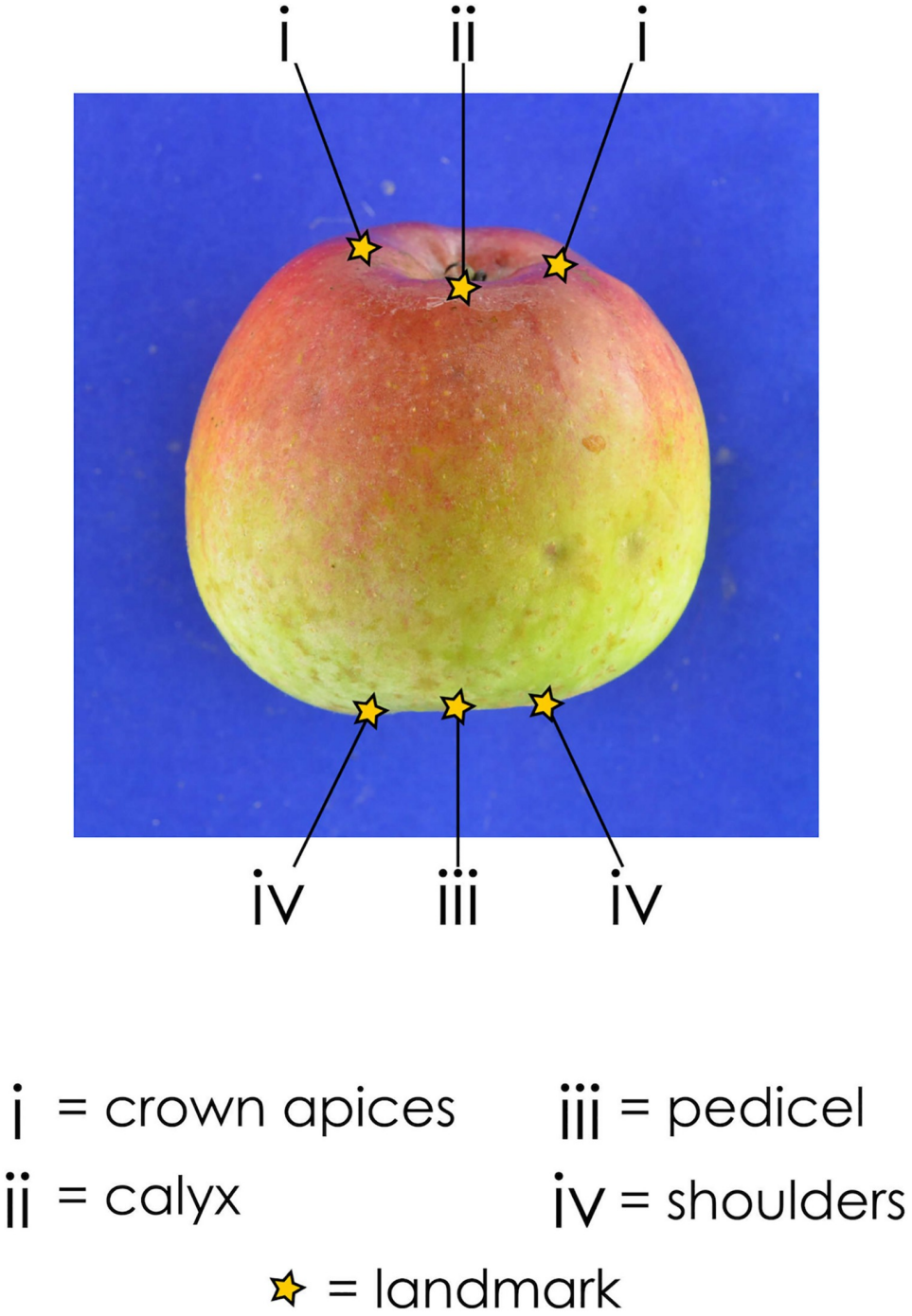

Fig 2. Selected landmarks for the geometric morphometrics dataset. Six landmarks were selected per image: two on the crown apices, one on the calyx, one on the pedicel attachment point and two on the shoulder apices.

https://doi.org/10.1371/journal.pone.0205357.g002

calyx area, and calyx ("eye") aperture area, and a geometric morphometrics dataset including: weight; first principal component of colour; calyx area; calyx ("eye") aperture area; Principal Component scores of Procrustes Coordinates, and centroid size. The datasets were then separated into training and testing sets with a 75–25% (15 fruit per cultivar in training, 5 in testing)

Table 1. Classifiers used to analyse linear and geometric morphometric datasets.

Classifier	Abbreviation
Adaptive Mixture Discriminant Analysis [47]	AMD
Bagged Classification and Regression Tree [48]	BCART
C5.0 Classification Tree [49]	C5.0
Classification and Regression Tree [50]	CART
Conditional Inference Random Forest [51]	CIRF
Feature Selection Random Forest [52]	FSRF
K-nearest Neighbor [53]	KNN
Naïve Bayes [54]	NB
Neural Network [55]	NN
Penalized Discriminant Analysis [56]	PDA
Robust Discriminant Analysis [57]	RDA
Random Ferns [58]	RF
Support vector Machine [59]	SVM

https://doi.org/10.1371/journal.pone.0205357.t001

[46] allocation respectively using identical partitions for comparability. The training sets were then used on 12 classifiers (Table 1). Using the same partition for both datasets ensured that the accuracy estimate for the test set of the best performing linear morphometrics classification was directly comparable to the accuracy estimate for the test set of the best performing geometric morphometrics classification.

Training used three repeats of 10-fold cross-validation. Each classifier was then tested using the test set and the classification confusion matrix, as well as the accuracy, kappa value, positive predictive rate, negative predictive rate, specificity, and sensitivity values were recorded. Paired t-tests on accuracy and kappa values were performed to compare between classifiers. The final model for each classification technique was selected in terms of highest accuracy and kappa values. Classification accuracy and kappa values using only colour are presented in S9 Table. All classification analysis was performed using the caret (Classification and Regression Training) package [60] in R [61].

To emulate the flexibility in character weighting shown by experts, who for instance might swap between using colour and size as a primary classifier, an ensemble approach was taken. When different datasets were used to train multiple classifiers, the success of each classifier with each cultivar could be recorded. For an unknown fruit tested against all the trained classifiers, the reliability of each prediction was assessed based on the accuracy of each classifier for the predicted cultivar. This process is illustrated in Fig 3. This replicated part of the expert flexibility by permitting the use of different characters for each classifier.

As an alternative approach to the manual ensemble procedure, the linear and geometric morphometrics datasets were combined to create a "kitchen sink" [62] dataset, to investigate whether the concatenation of raw data led to a more successful classification. The concatenated dataset was partitioned in the same way as the linear and geometric morphometrics datasets.

All the images used in the above study are deposited in the Reading Apple Image Library, accessible through the University of Reading Herbarium webpages. Together with the fully matured fruit presented in this study, the Image library also contains standardised images for fruit sampled from 12 of the cultivars at different time points from anthesis (weekly for the first two months from anthesis, and fortnightly later on). For each time-point, ten fruit were sampled and photographed as described here. Additionally, longitudinal sections from calyx to pedicel were performed and each side was photographed. For six cultivars, sampling was repeated for a second year. This resulted in 13360 images for 27 cultivars.

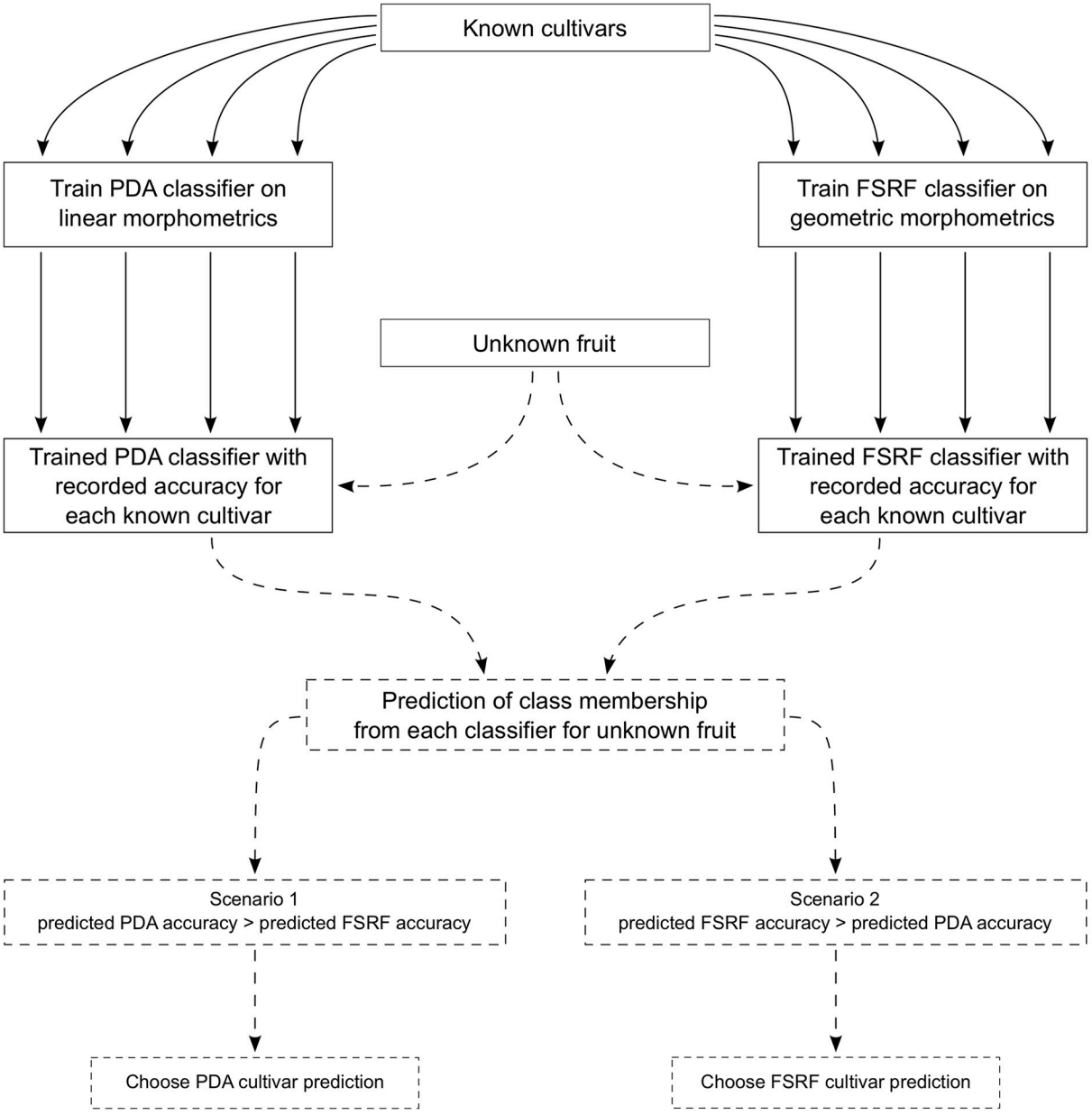

Fig 3. Flowchart of the manual ensemble process described in this work. After training the two classifiers and recording the cross-validation accuracy values for each cultivar, the unknown fruit is classified. The predicted class for the unknown fruit for each classifier is compared to the cross-validation accuracy for this class and the final prediction is selected according to which scenario applies.

https://doi.org/10.1371/journal.pone.0205357.g003

Results

Prior to training and testing of classifiers, the RGB colour values were reduced using principal component analysis (PCA). As the first principal component explained 94.3% of the overall colour variation, it was deemed a sufficient colour proxy, and was the only component retained for the remainder of the analysis. Classifier comparison was performed using accuracy and kappa values over the collected datasets. Classifiers were tested over four different settings: first against the linear morphometrics dataset, second against the geometric

morphometrics one, third using the manual ensemble approach, and finally against the kitchen-sink dataset. As benchmark, highest classification accuracy using colour alone was using a Support Vector Machine (accuracy: 27.4%, kappa:0.246) The remainder of this section is split to accommodate these four approaches.

Linear morphometrics

Of the 12 classifiers studied, Penalised Discriminant Analysis (PDA) had the highest mean accuracy and kappa values (accuracy: 73.0%, kappa: 0.722) for cross-validation of the training set and for this reason it was selected as the most appropriate classification technique. Following this, the test set, which comprised 135 fruit (5 from each cultivar) was analysed using the trained PDA classifier resulting in a percentage accuracy estimate over all classes (overall accuracy percentage) of 72.6%. Individual misclassifications for each fruit in the test set are in S4, S5 and S6 Tables.

Geometric morphometrics

Of the 11 classifiers tested, the best performing was the Feature Selection Random Forest (FSRF) as it had the highest mean accuracy and mean kappa values (accuracy: 66.5%, kappa: 0.654). Individual misclassifications for each fruit in the test set are in S7 and S8 Tables.

Manual ensemble

For manual ensemble, the predictions for the test set of the PDA on linear morphometrics were combined with the predictions of the FSRF classifier of the geometric morphometrics by using the accuracy estimates of cross-validation for each class (the detailed manual ensemble protocol is described in Materials and methods). The confusion matrix for the test set classification, which is the per-class performance of the trained classifier, is illustrated in Fig 4. Through the use of a heat-map, Fig 4 contrasts the actual class (cultivar) to which each fruit in the test set belonged (Reference) against what class it was predicted as (Prediction) by the trained classifier. Correct classifications are on the diagonal of the heat-map, with darker shades of blue illustrating greater success rates.

Kitchen-sink

Of the 11 classifiers tested with the "kitchen sink" dataset, the best performing was the Adaptive Mixture Discriminant Analysis (AMD) with mean accuracy of 70.5% and a kappa value of 0.692 for cross-validation.

The predictions of the test set samples for each classifier by cultivar are summarised in Fig 5, which demonstrates that every cultivar could be correctly classified using one of the four techniques. If one technique failed to classify a cultivar, another often turned out to be successful. The success of the classification techniques varied between the cultivars. For example, in the case of 'Adam's Pearmain' (Ada), all four approaches had a very high success rate, with three of them reaching 100% accuracy, and the lowest one reaching 80%. Findings were similar for 'Cloden' (Clo), with two methods reaching 100% accuracy and the remaining two 80%. Less successful was the case of 'Bovarde' (Bov), for which the FSRF classifier relying on the geometric morphometrics dataset failed to correctly identify any of the samples in the test set. Although none of the other classifiers succeeded in correctly identifying all five 'Bovarde' samples in the test set, they successfully identified three.

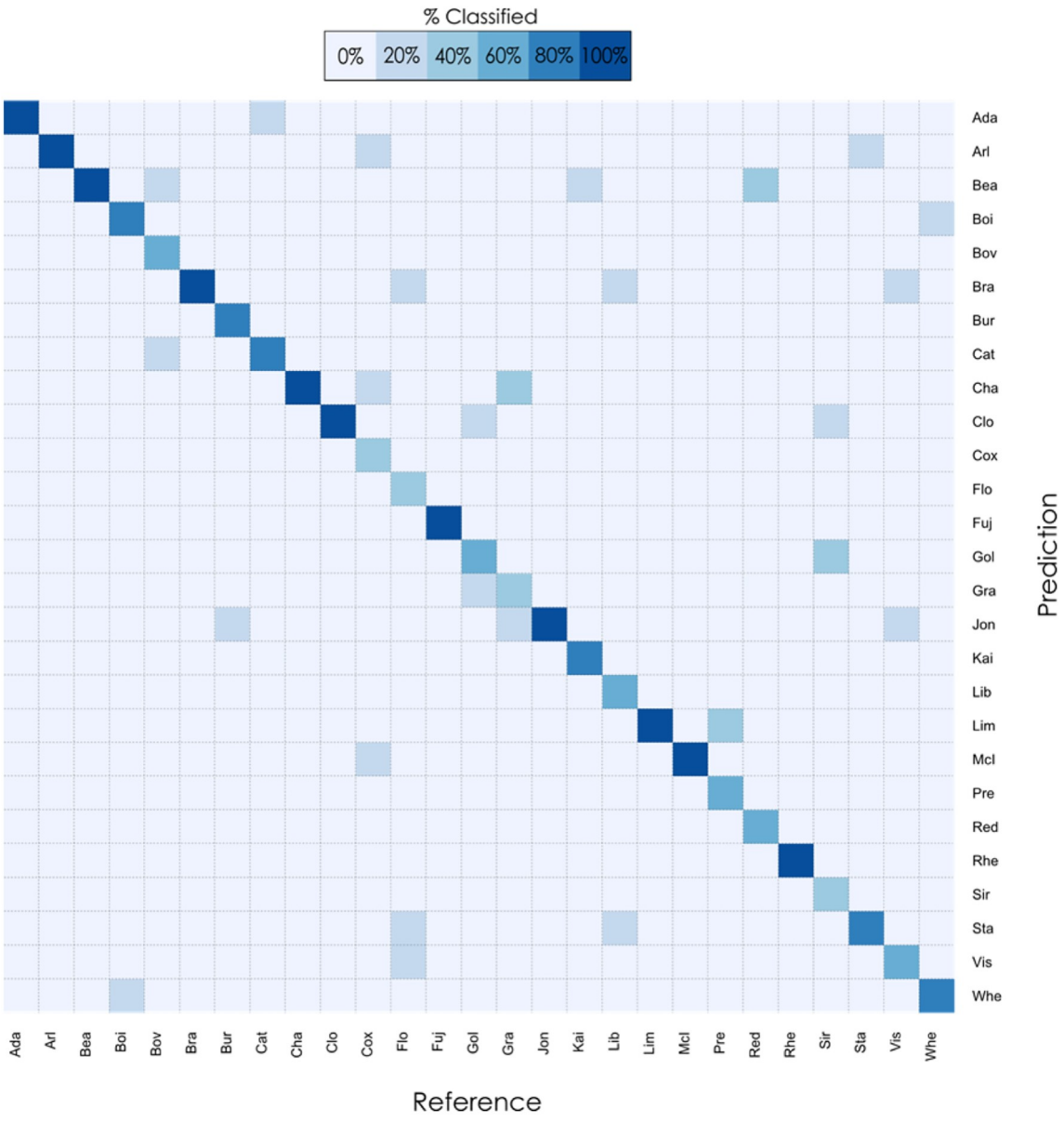

Fig 4. Confusion matrix from the Manual ensemble classification using the test set. The colours of the heat map correspond to the percentage of classification in each category. The accuracy obtained from the manual ensemble was 77.8% compared with 66.7% for the FSRF and 72.6% for the PDA on the same test set.

https://doi.org/10.1371/journal.pone.0205357.g004

Discussion

The advantage of studying apple cultivars which are clonally propagated was that we could be certain of the correct classification for each individual apple. This contrasts with equivalent studies of variation in species because species are conceptual constructs which may change over time [63,64]. For instance Compton and Hedderson [65] required 17 morphometric variables to separate a single variable species into four distinct ones, and those supported by correlation with geographic distribution. Despite the clonal identity within apple cultivars and the

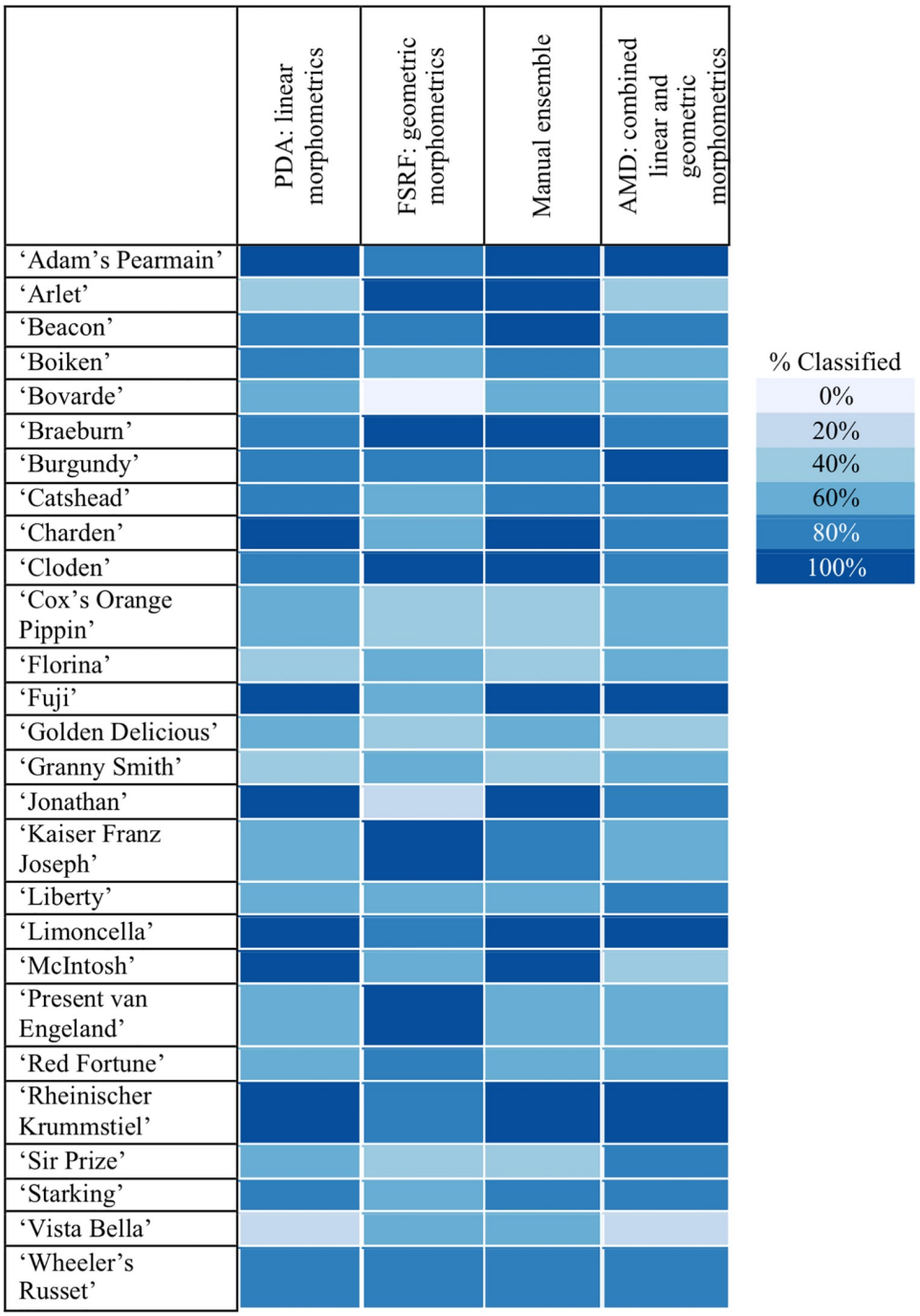

Fig 5. Summary of prediction rates of test set by cultivar for the classifiers using the same colours as the heatmap in Fig 3. Classifier abbreviations are explained in Materials and methods.

https://doi.org/10.1371/journal.pone.0205357.g005

variety of morphometric measurements used in this study, our classifications still resulted in misidentifications of many individual apples. Here we consider some of the underlying reasons for these.

We learned two major lessons during the process of automating classification.

Lesson 1: There is no free lunch

The performance and choice of classifier depends on the nature of the underlying data. For example, using linear morphometric techniques the best performing classifier was a PDA (accuracy 72.6%); for geometric morphometrics it was a FSRF (accuracy 66.7%). This finding is consistent with the "No free lunch" theorem. Stated formally by Wolpert and Macready [66], the theorem suggests that the performance of all classifiers is equal when the totality of possible problems is considered. This means that for every classifier there exists a possible problem where that classifier outperforms every other classifier. In our study two different morphometric datasets created two different classification problems, each analysed most effectively by a different classifier. This strong interaction between dataset and classifier is one of many examples of the no-free lunch theorem. Adding to the complexity is the impact of cultivar as a variable on the classifier and dataset interaction. As demonstrated in Fig 4, some cultivars were more accurately identified using one classifier and others by another. This suggests that in addition to selecting the appropriate classifier for the dataset, it is important to establish for every cultivar how accurately each combination performs. To illustrate this, four apples ('Arlet', 'Bovarde', 'Jonathan', 'Kaiser Franz Joseph') which were all part of the test set, are shown in Fig 6.

All of 'Arlet' (Arl) and 'Kaiser Franz Joseph'(Kai) samples in the test set were accurately classified using FSRF. Some 'Arlet' and 'Kaiser Franz Joseph' samples were misclassified using PDA (which had 40% success rate for 'Arlet' and 60% for 'Kaiser Franz Joseph'). 'Jonathan' (Jon) and 'Bovarde'(Bov) were classified more accurately by the PDA than the FSRF (100% and 60% respectively with the PDA as opposed to 20% and 0% with FSRF). Why are some cultivars more identifiable using one classifier than with another? For the cultivars that performed better with geometric morphometrics, such as 'Kaiser Franz Joseph', we propose that the distinctive fruit geometry failed to translate into recorded parameters in linear morphometrics. For the cultivars that performed better with the linear morphometrics, such as 'Jonathan', we propose that the overall geometry of the fruit was not as distinctive as the length and diameter measurements.

Lesson 2: Pick and mix

Improved accuracy results from the flexible combination of linear and geometric morphometrics classifiers. The successful protocol used as inspiration the flexibility of information that human identification experts can employ, by combining different data-sources (in this case linear and geometric morphometrics). The explanation for the superiority of this method (over both linear and geometric classifications) lies in the differences of accuracy per cultivar for each classification. The successful protocol gives different weights to the predictions depending on how accurate each classifier has been in the past for that particular prediction. For example, if an unknown fruit was predicted as 'Jonathan' by the FSRF and as 'McIntosh' by the PDA then the manual ensemble would classify it as a 'McIntosh' since the FSRF is weak at predicting 'Jonathan' (or 'McIntosh'), whereas the PDA is strong for both cultivars. By using this method and effectively relying on each classifier for the cultivars they were good at, the classification performance improved to an overall 77.8%. As a technique, it was particularly effective when there was a marked difference in the classification accuracy for a cultivar (e.g. with 'Jonathan'). When the classifiers performed at similar levels (e.g. 'Florina' with 40% with PDA and 60% with FSRF) then the approach chosen was sometimes the weaker one ('Florina' in manual ensemble had 40% accuracy). This was a result of the training accuracy being used to make a decision on the classifier for the test data in cases where training data did not support a clear decision.

Fig 6. Four fruit examples that were misclassified by one of the two classifiers. In the top two rows Arl and Kai were misclassified by the PDA but were successfully classified by the FSRF. In the bottom two rows, Bov and Jon were misclassified by the FSRF but successfully classified by the PDA.

https://doi.org/10.1371/journal.pone.0205357.g006

Although the "kitchen sink" approach was more accurate (71.1%) than the FSRF, it was less accurate than the PDA or manual ensemble. This indicates that the simple concatenation of both datasets increased noise. Aside from performance there was a fundamental difference between the "kitchen sink" and the manual ensemble. Both techniques used all the information available by including linear and geometric morphometrics but whereas the "kitchen sink" merged raw data, the manual ensemble exploited the strengths of each dataset.

Conclusions

The primary objective of this work was to discover whether apple cultivars could be identified using automated processes by exploiting some of the strategies apple experts employ in combination with current morphometric approaches. We conclude that computers can effectively simulate the approach used by apple experts, prioritising some data over others, in a cultivar- and situation-specific way.

The most impactful novelty of this work is methodological; specifically, the use of explicit geometric and linear morphometrics in combination with statistical learning has great relevance to wider biological research in identification and classification. It is not clear why such an ensemble method is not routinely used for biological identification as it combines the strength of several approaches. Ensemble learning techniques started gaining popularity in the 1990s for statistical learning specifically because they can combine weak learners (classifiers with low accuracy) to create a strong learner (classifier with high accuracy) [54]. Modern plant taxonomy could embrace this approach and take advantage of current computing power. This would permit the re-evaluation of data-sources which on their own may only lead to weak learners, but in thoughtful combinations have the potential to provide novel insight into classification of the organism under study. Crucially, the incorporation of multiple datasets towards a single classification problem is not about simply combining raw data from multiple sources; it is about the careful integration of such data and multiple approaches to analysis to improve insight and understanding.

Supporting information

S1 Fig. Accuracy and kappa value results with 95% confidence interval for the classification methods on the linear morphometrics dataset.
(DOCX)

S2 Fig. Accuracy and kappa value results with 95% confidence interval for the classification methods on the geometric morphometrics dataset.
(DOCX)

S1 Table. Selected cultivars for morphological classification study.
(DOCX)

S2 Table. Summary table of results from paired t-test comparisons between classifiers for linear morphometrics dataset. The results above the diagonal (represented by grey cells) were derived from t-tests on the accuracy values. The results below the diagonal were derived from t-tests on the kappa values. Significance is indicated with * levels (NS: Not Significant, *: $p \leq 0.05$, ** ≤ 0.01, *** ≤ 0.001).
(DOCX)

S3 Table. Summary table of results from paired t-test comparisons between classifiers for the geometric morphometrics dataset. The results above the diagonal (represented by grey cells) were derived from t-tests on the accuracy values. The results below the diagonal were derived from t-tests on the kappa values. Significance is indicated with * levels (NS: Not Significant, *:$p \leq 0.05$, ** ≤ 0.01, *** ≤ 0.001).
(DOCX)

S4 Table. Percentage match for misclassifications (Part 1 of 3) from linear morphometrics using Penalised Discriminant Analysis. All cultivars with at least a single misclassification are included. The highest three percentages for each cultivar are included. If the correct classification is not in the top three then it is also included together with its rank. If the fruit was correctly classified only the correct classification posterior is included.
(DOCX)

S5 Table. Percentage match for misclassifications (Part 2 of 3) from linear morphometrics using Penalised Discriminant Analysis. All cultivars with at least a single misclassification are included. The highest three percentages for each cultivar are included. If the correct

classification is not in the top three then it is also included together with its rank. If the fruit was correctly classified only the correct classification posterior is included.
(DOCX)

S6 Table. Percentage match for misclassifications (Part 3 of 3) from linear morphometrics using Penalised Discriminant Analysis. All cultivars with at least a single misclassification are included. The highest three percentages for each cultivar are included. If the correct classification is not in the top three then it is also included together with its rank. If the fruit was correctly classified only the correct classification posterior is included.
(DOCX)

S7 Table. Percentage match for misclassifications (Part 1 of 2) from geometric morphometrics using Feature Selection Random Forest. All cultivars with at least a single misclassification are included. The highest three percentages for each cultivar are included. If the correct classification is not in the top three then it is also included together with its rank. If the fruit was correctly classified only the correct classification posterior is included.
(DOCX)

S8 Table. Percentage match for misclassifications (Part 2 of 2) from geometric morphometrics using Feature Selection Random Forest. All cultivars with at least a single misclassification are included. The highest three percentages for each cultivar are included. If the correct classification is not in the top three then it is also included together with its rank. If the fruit was correctly classified only the correct classification posterior is included.
(DOCX)

S9 Table. Classification accuracy and kappa values for selected classifiers using only fruit colour. Classifier abbreviations follow Table 1.
(DOCX)

S1 File. Geometric morphometrics complete dataset.
(XLSX)

S2 File. Linear morphometrics complete dataset.
(XLSX)

Acknowledgments

We would like to thank Professor Richard Sibly, Dr Louise Johnson, Dr Tom Oliver, and Dr Jonathan Clark for constructive feedback on early drafts of this manuscript. We also thank Dr Matthew Ordidge for assistance with apple sampling, and Dr Kálmán Könyves for help with graphics.

Author Contributions

Conceptualization: Alastair Culham.

Data curation: Maria D. Christodoulou, Alastair Culham.

Formal analysis: Maria D. Christodoulou.

Funding acquisition: Nicholas Hugh Battey, Alastair Culham.

Methodology: Maria D. Christodoulou.

Project administration: Alastair Culham.

Supervision: Nicholas Hugh Battey, Alastair Culham.

Validation: Maria D. Christodoulou.

Writing – original draft: Maria D. Christodoulou, Alastair Culham.

Writing – review & editing: Maria D. Christodoulou, Nicholas Hugh Battey, Alastair Culham.

References

1. Noiton DAM, Alspach PA. Founding clones, inbreeding, coancestry, and status number of modern apple cultivars. J Am Soc Hortic Sci. American Society for Horticultural Science; 1996; 121: 773–782.
2. Qian G-Z, Liu L, Tang G. (1933) Proposal to conserve the name *Malus domestica* against *M. pumila*, *M. communis*, *M. frutescens*, and *Pyrus dioica* (Rosaceae). Taxon. 2010; 59: 3.
3. Lycett SJ, Von Cramon-Taubadel N. A 3D morphometric analysis of surface geometry in Levallois cores: Patterns of stability and variability across regions and their implications. J Archaeol Sci. Elsevier Ltd; 2013; 40: 1508–1517. https://doi.org/10.1016/j.jas.2012.11.005
4. Marramà G, Kriwet J. Principal component and discriminant analyses as powerful tools to support taxonomic identification and their use for functional and phylogenetic signal detection of isolated fossil shark teeth. PLoS One. 2017; 12: 1–22. https://doi.org/10.1371/journal.pone.0188806 PMID: 29182683
5. Rohlf F, Slice D. Extensions of the Procrustes method for the optimal superimposition of landmarks. Syst Biol. 1990; 39: 40–59. https://doi.org/10.2307/2992207
6. Rohlf FJ, Bookstein FL. Computing the uniform component of shape variation. Syst Biol. 2003; 52: 66–69. https://doi.org/10.1080/10635150390132759 PMID: 12554441
7. Adams DC, Rohlf FJ, Slice DE. Geometric morphometrics: Ten years of progress following the 'revolution'. Ital J Zool. 2004; 71: 5–16. https://doi.org/10.1080/11250000409356545
8. Stoyanova DK, Algee-Hewitt BFB, Kim J, Slice DE. A Computational Framework for Age-at-Death Estimation from the Skeleton: Surface and Outline Analysis of 3D Laser Scans of the Adult Pubic Symphysis. J Forensic Sci. 2017; 62: 1434–1444. https://doi.org/10.1111/1556-4029.13439 PMID: 28244105
9. Linnaeus C. Species plantarum: A facsimile of the first edition. London: The Ray Society 1957; 1753.
10. Palmeri TJ, Gauthier I. Visual object understanding. Nat Rev Neurosci. 2004; 5: 291–303. https://doi.org/10.1038/nrn1364 PMID: 15034554
11. von Ahn L, Blum M, Hopper NJ, Langford J. CAPTCHA: Using hard AI problems for security. Advances in Cryptology—EUROCRYPT 2003. 2003. pp. 294–311. https://doi.org/10.1007/3-540-39200-9_18
12. Zhao W, Chellappa R, Phillips PJ, Rosenfeld A. Face recognition: A literature survey. ACM Comput Surv. 2003; 35: 399–458. https://doi.org/10.1145/954339.954342
13. Liu J, Li J, Feng L, Li L, Tian J, Lee K. Seeing Jesus in toast: Neural and behavioral correlates of face pareidolia. Cortex. Elsevier Ltd; 2014; 53: 60–77. https://doi.org/10.1016/j.cortex.2014.01.013 PMID: 24583223
14. Sinha P. Recognizing complex patterns. Nat Neurosci. 2002; 5 Suppl: 1093–7. https://doi.org/10.1038/nn949 PMID: 12403994
15. Zheng L, He X, Hintz T. Comparison of SVMs in number plate recognition. In: Singh S, Singh M, editors. Progress in Pattern Recognition. London: Springer-Verlag; 2007.
16. Van Bocxlaer B, Schultheiß R. Comparison of morphometric techniques for shapes with few homologous landmarks based on machine-learning approaches to biological discrimination. Paleobiology. 2010; 36: 497–515. https://doi.org/10.1666/08068.1
17. Guisande C, Manjarrés-Hernández A, Pelayo-Villamil P, Granado-Lorencio C, Riveiro I, Acuña A, et al. IPez: An expert system for the taxonomic identification of fishes based on machine learning techniques. Fish Res. 2010; 102: 240–247. https://doi.org/10.1016/j.fishres.2009.12.003
18. Santana FS, Costa AHR, Truzzi FS, Silva FL, Santos SL, Francoy TM, et al. A reference process for automating bee species identification based on wing images and digital image processing. Ecol Inform. Elsevier B.V.; 2014; 24: 248–260. https://doi.org/10.1016/j.ecoinf.2013.12.001
19. da Silva FL, Sella MLG, Francoy TM, Costa AHR. Evaluating classification and feature selection techniques for honeybee subspecies identification using wing images. Comput Electron Agric. Elsevier B. V.; 2015; 114: 68–77. https://doi.org/10.1016/j.compag.2015.03.012
20. Velemínská J, Krajíček V, Dupej J, Goméz-Valdés JA, Velemínský P, Šefčáková A, et al. Technical Note: Geometric morphometrics and sexual dimorphism of the greater sciatic notch in adults from two skeletal collections: The accuracy and reliability of sex classification. Am J Phys Anthropol. 2013; 152: 558–565. https://doi.org/10.1002/ajpa.22373 PMID: 24114412

21. Wilf P, Zhang S, Chikkerur S, Little SA, Wing SL, Serre T. Computer vision cracks the leaf code. Proc Natl Acad Sci. 2016; 201524473. https://doi.org/10.1073/pnas.1524473113 PMID: 26951664

22. DEFRA. Wholesale fruit and vegetable prices, weekly average [Internet]. 2018.

23. Pante E, Schoelinck C, Puillandre N. From integrative taxonomy to species description: One step beyond. Syst Biol. 2015; 64: 152–160. https://doi.org/10.1093/sysbio/syu083 PMID: 25358968

24. Mckay BD, Mays HL, Yao C Te, Wan D, Higuchi H, Nishiumi I. Incorporating color into integrative taxonomy: Analysis of the varied tit (*Sittiparus varius*) complex in East Asia. Syst Biol. 2014; 63: 505–517. https://doi.org/10.1093/sysbio/syu016 PMID: 24603127

25. Magauer M, Schönswetter P, Jang TS, Frajman B. Disentangling relationships within the disjunctly distributed *Alyssum ovirense*/*A.wulfenianum* group (Brassicaceae), including description of a novel species from the north-eastern Alps. Bot J Linn Soc. 2014; 176: 486–505. https://doi.org/10.1111/boj.12214

26. Lecocq T, Dellicour S, Michez D, Dehon M, Dewulf A, De Meulemeester T, et al. Methods for species delimitation in bumblebees (Hymenoptera, Apidae, *Bombus*): Towards an integrative approach. Zool Scr. 2015; 44: 281–297. https://doi.org/10.1111/zsc.12107

27. Ronikier M, Zalewska-Gałosz J. Independent evolutionary history between the Balkan ranges and more northerly mountains in *Campanula alpina* s.l. (Campanulaceae): Genetic divergence and morphological segregation of taxa. Taxon. 2014; 63: 116–131. https://doi.org/10.12705/631.4

28. Schmidt-Roach S, Miller KJ, Lundgren P, Andreakis N. With eyes wide open: A revision of species within and closely related to the *Pocillopora damicornis* species complex (Scleractinia; Pocilloporidae) using morphology and genetics. Zool J Linn Soc. 2014; 170: 1–33. https://doi.org/10.1111/zoj.12092

29. Mercês MP, Lynch Alfaro JW, Ferreira WAS, Harada ML, Silva Júnior JS. Morphology and mitochondrial phylogenetics reveal that the Amazon River separates two eastern squirrel monkey species: *Saimiri sciureus* and *S. collinsi*. Mol Phylogenet Evol. 2015; 82: 426–435. https://doi.org/10.1016/j.ympev.2014.09.020 PMID: 25451802

30. Skoracka A, Kuczyński L, Rector B, Amrine JW. Wheat curl mite and dry bulb mite: Untangling a taxonomic conundrum through a multidisciplinary approach. Biol J Linn Soc. 2014; 111: 421–436. https://doi.org/10.1111/bij.12213

31. Mamos T, Wattier R, Majda A, Sket B, Grabowski M. Morphological vs. Molecular delineation of taxa across montane regions in Europe: The case study of *Gammarus balcanicus* Schäferna, (Crustacea: Amphipoda). J Zool Syst Evol Res. 2014; 52: 237–248. https://doi.org/10.1111/jzs.12062

32. Laurito M, Almirón WR, Ludueña-Almeida FF. Discrimination of four *Culex* (*Culex*) species from the Neotropics based on geometric morphometrics. Zoomorphology. 2015; 134: 447–455. https://doi.org/10.1007/s00435-015-0271-x

33. Buj I, Šanda R, Marčić Z, Ćaleta M, Mrakovčić M. Combining morphology and genetics in resolving taxonomy-a systematic revision of spined loaches (genus *Cobitis*; Cypriniformes, Actinopterygii) in the adriatic watershed. PLoS One. 2014; 9. https://doi.org/10.1371/journal.pone.0099833 PMID: 24918426

34. Clark S, Cleal Q. A manual key for the identification of apples based on descriptions in Bultitude (1983). Yorkshire; 2005.

35. Sanders R. The Apple Book. 1st ed. London: Frances Lincoln Limited Publishers; 2010.

36. Morgan J, Richards A. The Book of Apples. 1st ed. London: Ebury Press; 1993.

37. Angelova A, Zhu S. Efficient object detection and segmentation for fine-grained recognition. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit. 2013; 811–818. https://doi.org/10.1109/CVPR.2013.110

38. Nilsback ME, Zisserman A. Automated flower classification over a large number of classes. Proc - 6th Indian Conf Comput Vision, Graph Image Process ICVGIP 2008. 2008; 722–729. 10.1109/ICVGIP.2008.47

39. Poland J, Clement EJ. The Vegetative Key to the British Flora. Botanical Society of the British Isles; 2009.

40. Corney DPA, Tang HL, Clark JY, Hu Y, Jin J. Automating digital leaf measurement: The tooth, the whole tooth, and nothing but the tooth. PLoS One. 2012; 7: 1–10. https://doi.org/10.1371/journal.pone.0042112 PMID: 22870286

41. Corney DPA, Clark JY, Tang HL, Wilkin P. Automatic extraction of leaf characters from herbarium specimens. Taxon. 2012; 61: 231–244.

42. Clark JY, Corney DPA, Wilkin P. Leaf-based automated species classification using image processing and neural networks. In: Lestrel PE, editor. Proceedings of the 4th International Symposium on Biological Shape Analysis (ISBSA). World Scientific; 2017. pp. 29–56.

43. Rohlf FJ. tpsDig 2.17 [Internet]. Stony Brook; 2013. http://life.bio.sunysb.edu/morph/soft-dataacq.html

44. Klingenberg C. Morpho J: an integrated software package for geometric morphometrics. Mol Ecol Resour. 2011; 11: 353–7. https://doi.org/10.1111/j.1755-0998.2010.02924.x PMID: 21429143

45. Abràmoff MD, Magalhães PJ, Ram SJ. Image processing with ImageJ. Biophotonics Int. 2004; 11: 36–41.

46. Kampichler C, Wieland R, Calmé S, Weissenberger H, Arriaga-Weiss S. Classification in conservation biology: A comparison of five machine-learning methods. Ecol Inform. 2010; 5: 441–450. https://doi.org/10.1016/j.ecoinf.2010.06.003

47. Bouveyron C. Adaptive mixture discriminant analysis for supervised leaning with unobserved classes. J Classif. 2014; 31: 49–84.

48. Breiman L. Bagging predictors. Mach Learn. 1996; 24: 123–140.

49. Quinlan JR. C5.0 version 2.07 [Internet]. Empire Bay, Australia; 2015 [cited 16 Sep 2015]. http://www.rulequest.com/download.html

50. Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and Regression Trees. London: Chapman & Hall; 1984.

51. Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: A conditional inference framework. J Comput Graph Stat. 2006; 15: 651–674. https://doi.org/10.1198/106186006X133933

52. Breiman L. Random forests. Mach Learn. 2001; 45: 5–32. https://doi.org/10.1023/A:1010933404324

53. Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans Inf Theory. 1967; 13: 21–27. https://doi.org/10.1109/TIT.1967.1053964

54. John GHG, Langley P. Estimating continuous distributions in Bayesian classifiers. Eleventh Conference on Uncertainty in Artificial Intelligence. 1995. pp. 338–345. 10.1.1.8.3257

55. Werbos P. Beyond regression: New Tools for prediction and analysis in the behavioral sciences. Harvard University. 1974.

56. Hastie T, Buja A, Tibshirani R. Penalized discriminant analysis. Ann Stat. 1995; 23: 73–102.

57. Kim SJ, Magnani A, Boyd SP. Robust Fisher discriminant analysis. Electr Eng. 2006; 1: 1–8.

58. Ozuysal M, Calonder M, Lepetit V, Fua P. Fast keypoint recognition using random ferns. IEEE Trans Pattern Anal Mach Intell. 2010; 32: 448–461. https://doi.org/10.1109/TPAMI.2009.23 PMID: 20075471

59. Cortes C, Vapnik V. Support vector network. Mach Learn. 1995; 20: 1–25.

60. Kuhn M, Wing J, Weston S, Williams A, Keefer C, Engelwhardt A, et al. caret: Classification and Regression Training. R package version 6.0–37. 2014.

61. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. [Internet]. Vienna, Austria.: R Foundation for Statistical Computing; 2017. http://www.r-project.org/.

62. Christodoulou MD. Quantification of fruit shape in apple: Development of methodologies and assessment of their potential use in cultivar identification. 2016.

63. Hey J, Waples RS, Arnold ML, Butlin RK, Harrison RG. Understanding and confronting species uncertainty in biology and conservation. Trends Ecol Evol. 2003; 18: 597–603. https://doi.org/10.1016/j.tree.2003.08.014

64. Hey J. On the failure of modern species concepts. Trends Ecol Evol. 2006; 21: 447–450. https://doi.org/10.1016/j.tree.2006.05.011 PMID: 16762447

65. Compton JA, Hedderson TA. A morphometric analysis of the Cimicifuga foetida L. complex (Ranunculaceae). Bot J Linn Soc. 1997; 123: 1–23. https://doi.org/10.1111/j.1095-8339.1997.tb01402.x

66. Wolpert DH, Macready WG. No free lunch theorems for optimization. IEEE Trans Evol Comput. 1997; 1: 67–82. https://doi.org/10.1109/4235.585893