Palaeontology
The Palaeontological Association
Registered Charity No. 11683330

Learning to see the wood for the trees: machine learning, decision trees and the classification of isolated theropod teeth

Journal:	*Palaeontology*
Manuscript ID	PALA-05-20-4785-OA.R1
Manuscript Type:	Original Article
Date Submitted by the Author:	09-Aug-2020
Complete List of Authors:	Wills, Simon; Natural History Museum, Earth Sciences Underwood, Charlie; Birkbeck College, Department of Earth and Planetary Sciences Barrett, Paul; Natural History Museum, Earth Sciences, Cromwell Road, South Kensington
Key words:	machine learning, decision trees, classification, random forests, theropod teeth, discriminant analysis

LEARNING TO SEE THE WOOD FOR THE TREES: MACHINE LEARNING, DECISION TREES AND THE CLASSIFICATION OF ISOLATED THEROPOD TEETH

SIMON WILLS[1,2]*, CHARLIE J. UNDERWOOD[2] *and* PAUL M. BARRETT[1]

[1] Department of Earth Sciences, Natural History Museum, Cromwell Road, South Kensington, London SW7 5BD, United Kingdom; s.wills@nhm.ac.uk

[2] Department of Earth and Planetary Sciences, Birkbeck College, Malet Street, London WC1E 7HX, United Kingdom

* Corresponding author

Abstract: Taxonomic identification of fossils based on morphometric data traditionally relies on the use of standard linear models to classify such data. Machine learning and decision trees offer powerful alternative approaches to this problem but are not widely used in palaeontology. Here, we apply these techniques to published morphometric data of isolated theropod teeth in order to explore their utility in tackling taxonomic problems. We chose two published datasets consisting of 886 teeth from 14 taxa and 3020 teeth from 17 taxa, respectively, each with five morphometric variables per tooth. We also explored the effects that missing data have on the final classification accuracy. Our results suggest that machine learning and decision trees yield superior classification results over a wide range of data permutations, with decision trees achieving accuracies of 96% in classifying test data in some cases. Missing data or attempts to generate synthetic data to overcome missing data seriously degrade all classifiers predictive accuracy. The results of our analyses also indicate that using ensemble classifiers combining different classification techniques and the examination of posterior probabilities is a useful aid in checking final class assignments. The application of such techniques to isolated theropod teeth demonstrate that simple morphometric data can be used to yield statistically robust taxonomic classifications and that lower classification accuracy is more likely to reflect preservational limitations of the data or poor application of the methods.

Key words: machine learning, discriminant analysis, decision trees, classification, Theropoda, teeth.

The use of non-linear analytical techniques (Table 1) that draw upon the rapidly expanding field of machine learning and decision trees has remained mostly unexplored with respect to characterizing fossil vertebrate morphology (Monson *et al.* 2018). By contrast, other disciplines have rapidly embraced machine learning techniques to undertake classification, prediction and various modelling tasks (Christin *et al.* 2019). Applications range from ecological modelling (Džeroski 2001; Cutler *et al.* 2007), population monitoring (Britzke *et al.* 2011), automated taxonomic classification by phenotype (Hoyal Cuthill *et al.* 2019), medical image analysis (Ker *et al.* 2018), financial modelling and prediction (De Spiegeleer *et al.* 2018; Ma and Lv 2019), psychology (Holden *et al.* 2011; Finch *et al.* 2014) and bioinformatics (Chen and Ishwaran 2012; Couronné *et al.* 2018) to the digitisation of natural history collections (Schuettpelz *et al.* 2017). Automated and semi-automated approaches of data modelling have also been used for taxon identification and dietary inference from tooth surface morphology (Evans *et al.* 2007; MacLeod 2007, 2015, 2017; Wilson *et al.* 2012; Melstrom and Irmis 2019) and are commony used in the analysis of earth observation data (Onojeghuo *et al.* 2018; Son *et al.* 2018; MacLeod 2019).

Here we test the suitability of these methods for the taxonomic identificiation of fossils, using isolated non-avian theropod dinosaur teeth as a case study. Previously, standard linear classification models have been used to classify these specimens based on shape data (see below). Here we apply several alternative approaches to this problem and assess their comparative performance based on analysis of two datasets of isolated theropod tooth measurements.

The regular shedding of functional teeth (Currie *et al.* 1990; Farlow *et al.* 1991), plus their resistance to abrasion and chemical alteration (Argast *et al.* 1987), results in the recovery of abundant, isolated dinosaur teeth in many Mesozoic terrestrial deposits (e.g., Evans and Milner 1994; Fiorillo and Currie 1994; Metcalf and Walker 1994; Rauhut 2002; Sankey *et al.* 2002; Knoll and Ruiz-Omeñaca 2009; Larson and Currie 2013; Gates *et al.* 2015). These teeth represent the vast majority of dinosaur material recovered from microvertebrate localities, and often represent the only source of information for interpretations of dinosaur species-richness and palaeoecology from such sites (e.g., Williamson and Brusatte 2014; Wings *et al.* 2015; Larson *et al.* 2016). A reliable, repeatable framework for assessing the taxonomic identity of isolated teeth would therefore be useful in providing more accurate assessments of the faunal compositions of both microvertebrate localities and other localities where skeletal material is rare or uncommon. Historically, three positions have been taken on the taxonomic utility of isolated dinosaur teeth (Heckert 2002): (1) that teeth are almost entirely non-diagnostic at generic or specific level and have little or no taxonomic value (e.g., Charig and Crompton 1974; Ostrom and Wellnhofer 1990; Dodson and Dawson 1991); (2) that teeth have some diagnostic value, but in the absence of other skeletal material the use of isolated teeth in diagnosing taxa to higher taxonomic levels is questionable (e.g., Currie*, et al.* 1990; Padian 1990; Sereno 1991; Larson and Currie 2013); and (3) that dinosaur teeth can be taxonomically diagnostic and bear synapomorphies that can be used to erect valid taxa or assign isolated teeth to known existing taxa (e.g., Thulborn 1973, 1992; Hunt and Lucas 1994; Heckert 2002, 2004; Hendrickx *et al*. 2020). Recent work based on detailed character descriptions, morphometric analyses, or a combination of these approaches indicates that at least some diagnostic value can be extracted from dinosaur teeth (e.g., Smith 2005; Smith *et al.* 2005; Larson and Currie 2013; Barrett *et al.* 2014; Hendrickx and Mateus 2014; Boyd 2015; Hendrickx *et al.* 2015, 2019; Ősi *et al.* 2016; Strickson *et al.* 2016). Nevertheless, as tooth morphology can vary ontogenetically, positionally (within the jaws of the same animal) and between individuals, as well as taxonomically (Coombs 1990; Hendrickx *et al*. 2019), there is still disagreement regarding the most appropriate method for assigning isolated teeth to defensible, recognizable morphotaxa, which could then form a basis for further investigation. Indeed, Hendrickx *et al.* (2015, 2020) have suggested that morphometric data alone are sub-optimal for classification and that far better results can be obtained using detailed descriptions of morphological characters and cladistic analyses based on a dentition-based data matrix.

Currie *et al.* (1990) and Farlow *et al.* (1991) were the first to apply a morphometric approach to isolated dinosaur teeth in a systematic fashion to aid taxonomic identification and examine the functional significance of different tooth crown morphologies. Smith (2005) and Smith *et al.* (2005), building on previous work (e.g., Chandler 1990; Currie *et al.* 1990; Farlow *et al.* 1991; Baszio 1997), provided a preliminary framework for the taxonomic identification of theropod dinosaur teeth by applying multivariate statistical methods to standard morphometric measurements. Following this work a generic approach applying principle component analysis (PCA) and linear discriminant analysis (LDA) has become the ‘standard’ quantitative methodology for the identification of isolated theropod teeth (e.g., Samman *et al.* 2005; Fanti and Therrien 2007; Larson 2008; Larson and Currie 2013; Williamson and Brusatte 2014; Torices *et al.* 2014; Hendrickx *et al.* 2015; Gerke and Wings 2016; Young *et al.* 2019). Similar methodologies have been applied to ornithischian dinosaurs (Becerra *et al.* 2013) and isolated teeth from other extinct taxa, such as sharks (Marramà and Kriwet 2017) and archosauriforms (Hoffman *et al.* 2019).

However, caution is warranted when applying this methodology. The use of PCA alone is not suitable to assess between-group differences and can mask differences when the group structure is embedded within variables exhibiting lower variances (MacLeod 2018), or when group differences are assessed on a limited number of principle components by simply plotting PC1 against PC2. It is, however, useful as a dimensionality reduction transformation where there is a requirement to reduce the number of predictor variables while retaining the embedded information content, or as an investigative tool to explore data structures (Jolliffe 2002; MacLeod 2018). LDA is commonly used as either a follow-on classifier from PCA – by submitting the retained PCA eigenvectors to the LDA model – or as a classifier applied directly to the raw data. Most applications of LDA assume that the data under investigation meets the requirements of the technique, but do not always check that this is the case. This is important, as LDA can be adversely affected by small or widely unequal group sizes, data outliers, unequal covariance matrices and non-Gaussian distributions, and the method works more effectively when the smallest group has significantly more cases than predictor variables. The effects of these caveats may be marginal in practice (Feldesman 2002) but thus far these issues have not received detailed discussion in this context. If the data under consideration do violate these assumptions it calls into question the results obtained from such analyses, especially in the absence of verification by other methods (e.g., Whitenack and Gottfried 2010; Fraser and Theodor 2011; Hendrickx *et al*. 2015, 2019; Milla Carmona *et al*. 2016; Corentin and Salvador 2018).

The algorithms employed in these analyses (Table 2) belong to a category of supervised classifiers known as 'eager-learners', where a model is generated from a set of training data before being applied to an 'unknown' dataset. The function of a supervised classifier is to build a model that then enables correct assignment of a future object described by predictor variables to a known class (Rausch and Kelley 2009; Maugis *et al.* 2011). Eager-learners often take a long time to construct a model but can make predictions quickly. It is also possible to use some of these techniques, such as random forests, in unsupervised mode to assess and detect meaningful structures in a dataset and to classify objects to groups that are not known *a priori* (Shi and Horvath 2006; Criminisi *et al.* 2012; Afanador *et al.* 2016). Although we have employed these techniques on fairly simple morphometric measurements, there is no reason why the techniques discussed below could not be employed on more complex morphological datasets such as 3D-shape data or digital images. Below we include a short introduction to the techniques we applied, including the use of ensemble model classifiers.

Linear models

Linear discriminant analysis. Linear discriminant analysis (LDA), a technique that identifies linear combinations of predictor variables to maximise the multivariate distance between groups (Fisher 1936; Welch 1939), is perhaps the most widely used method for classification. The functions are calculated in such a way that the first function captures as much of the group differences as possible, with subsequent functions each representing group differences not captured by previous functions. The combinations of predictors and prior probabilities are then used to calculate the posterior probability distribution for each case. Group membership is assigned by selecting the group with the highest posterior probability for each case. For LDA to function appropriately two underlying assumptions regarding the data are made: that the data is multivariate normal; and that the group covariance matrices for the predictor variables are equal (Feldesman 2002; Hastie *et al.* 2009a). LDA is also sensitive to highly-correlated predictors and is dependent on the ability to invert the covariance matrix, requiring more samples than predictors per group.

Logistic Regression. Logistic regression (LR), although commonly used to solve two-class problems, can be extended to a multi-class scenario and uses a linear predictor function to assess the likelihood of a particular class outcome. LR uses the log of the odds of being in one group compared to the others as the basis of its prediction. No assumptions are made regarding the distribution of the predictor variables entered into the model, nor does it assume equal covariance matrices and therefore no additional data pre-processing is required (Rausch and Kelley 2009; Kuhn and Johnson 2013a; Finch *et al.* 2014).

Non-linear models

Mixture Discriminant Analysis. Mixture discriminant analysis (MDA) is a non-linear extension of LDA whereby each class is modelled as a mixture of multiple multivariate normal distributions, i.e., each class can contain an unobserved number of sub-classes (Hastie and Tibshirani 1996; Kuhn and Johnson 2013a; Finch *et al.* 2014). Unlike LDA, there is no assumption of equal covariance matrices across groups for MDA. In a biological classification of taxa such sub-classes are particularly relevant, especially when classifying data to higher taxonomic levels. MDA has been applied with some success in other fields and often exhibits high predictive accuracy (Rausch and Kelley 2009; Britzke *et al.* 2011; Finch *et al.* 2014).

Naïve Bayes. Naïve Bayes (NB) is a non-linear machine learning approach to group classification (Russell and Norvig 2009; Marsland 2015) that is known to work well with small group sample sizes (MacLeod 2018). The model assumes that all the predictors are independent of each other which results in relatively quick computational times (Kuhn and Johnson 2013a).

Decision trees

The final methodologies we explore are a departure from the standard linear or non-linear families of classification models. Both random forests and C5.0 are decision tree-based techniques that expand on the seminal work of Breiman *et al.* (1984), which introduced classification and regression trees.

Before exploring the detail of the two techniques it is useful to understand the basics of a decision tree. Decision trees are used in everyday life to make decisions based on a series of criteria. A simple example would be to decide on which train to catch to reach a certain destination at a preferred time without changing stations. In order to reach this decision we effectively run through a series of steps, each step is a question and the answer to the question dictates a path that the decision can follow. A suitable decision tree for such a choice is shown in Figure 1. Every decision tree is a nested hierarchy of questions and answers (or if/then statements). For the example of catching a train to London Victoria station, the following hypothetical decision tree (one of many possible trees) might be followed:

If the final destination of the train is Brighton, then it is the wrong train

or

If the final destination of the train is London, and the station is London Bridge, then it is the wrong train

or

If the final destination of the train is London, and the station is Victoria, and it is a not direct train, then it is the wrong train

or

If then final destination of the train is London, and the station is Victoria, and it is a direct train, and it arrives between 08:00 and 08:15, then it is the correct train.

A decision tree is essentially a flowchart of questions or rules that leads down a path to a prediction. Data is inputted into the root node of the tree. The decision tree algorithm then progressively divides the data into smaller and smaller groups based on the splitting criteria until the point at which the dataset can either be split no more or it reaches a rule that orders the splitting to finish. Decision trees can either be regression trees where the predicted outcome is a value (e.g., a house price) or classification trees where the predicted outcome is categorical (e.g., a taxon). The concepts of decision trees and random forests are similar. A decision tree is effectively built upon the entire dataset to produce one tree. A random forest combines many decision trees into a single model, where each of the trees in the model is generated on random subsets of observations and variables.

The major advantages of decision trees over techniques such as LDA or logistic regression are that: 1) they can accommodate missing data; 2) there is no need for the data to conform to a normal distribution, as the techniques are non-parametric; 3) outliers have little effect on the final classification as they will rarely define a splitting node; 4) they can use both categorical and numerical data as predictor variables; and 5) transformed predictor variables (e.g., log transforms) have no effect on the tree structure (Feldesman 2002). A drawback with decision tree methods is that of overfitting the data. This occurs when a tree is grown that perfectly predicts the classification pattern of the training data by defining terminal nodes (or leaves) that fit particular idiosyncrasies of the training process, i.e. that are relevant to that particular dataset only. Tree-based methods are also prone to bias if some classes dominate the data and care needs to be taken to account for this prior to fitting.

Random Forests. Random forests (RF) is an ensemble learning method where a large number of uncorrelated decision trees are aggregated to form a final classification (Breiman 2001). This final classification is based on either an average of all the individual tree estimates (for regression trees) or a simple majority vote (for classification trees). The decision trees are built by randomly selecting predictors and observations to create individual trees. This random selection process increases the diversity in the forest and leads to a more robust prediction. Random predictors (i.e., variables) are used at each split in the tree which de-correlate the trees forming the forest. The number of predictors used is controlled by a parameter setting (m_{try}) which Kuhn and Johnson (2013a) and Breiman (2001) recommend setting to the square root of the number of predictors. RF classifications are sensitive to the number of trees used to build the forest with error rates reducing with increasing numbers of trees. Random forests tend to be stable and produce good predictive performance. However, they do have a number of disadvantages: even though some parameters are controllable, such as the number of trees or the number of predictors available at each split, the actual make up of each tree and therefore the forest is random and the forest itself (not the prediction) is less easy to interpret than a single decision tree; training a large number of trees can have higher computational overhead than a simple single decision tree.

C5.0. The C5.0 rule-based decision tree classifier is an updated version of the C4.5 model of Quinlan (1993) where the splitting criterion is based on information theory to choose the most informative variables for classifying the training set (Kuhn and Johnson 2013a; Mehta and Shukla 2015). As with decision trees in general, each sub-set resulting from the initial split is then re-split (usually on a different field) with the process repeating until no more splitting is possible. Each split can be either binary or multi-branched. C5.0 then tries to reduce the effects of overfitting by undertaking a pruning (winnowing) process on the lower level splits to remove those that do not contribute to the final model and produce simpler and more accurate trees. Unlike random forests the C5.0 tree is built by default on the entire dataset using all the variables and cases. The winnowing process attempts to uncover predictor variables that have a relationship to the desired model outcome with the final model only built using those variables. The C5.0 algorithm also allows for the implementation of adaptive boosting, which generates multiple classifiers rather than one with the final prediction resulting from majority voting across the classifiers. Unlike random forests, which creates multiple random trees, the C5.0 adaptive boosting trees are linked back to the classification errors generated from the first tree or ruleset. The first classifier will usually make mistakes on some groups. A second classifier is then generated that focusses on the misclassified data from the first tree in an attempt to improve the misclassification rate. Errors from the second tree are passed to a third and so on. The process continues for a user pre-defined number of iterations (trials). For a more detailed description of both C4.5 and C5.0 methods see Kuhn and Johnson (2013a).

Ensemble models

Ensemble learning methods take a series of classifier models and combine the predictions to produce a final classification (Dietterich 2001; Roli *et al.* 2001). A key to a good ensemble model is that the individual classifier techniques should be diverse to create a stronger overall prediction. There are a number of different methods to combine the results of the models making up the ensemble such as bagging, boosting and stacking (Dietterich 2001), here we use majority-voting and model stacking to arrive at the final classification. Majority voting simply takes the majority rule of the predictions from each classifier as the final classification result, for example if two classifiers predict a case to be ‘class 1’ and one classifier predicts the case as ‘class 2’ then the ensemble classification for that case is ‘class 1’. Model stacking is where a single training dataset is run through multiple models. The predictions from these models are then used as the input to a second-level model from which the final classification is drawn.

MATERIALS AND METHODS

Here we describe the datasets used for the analysis and the data preparation steps involved. We also discuss how we dealt with common issues found in multivariate datasets used for classification models, such as class balancing and missing data. In addition, we examine how the choice of prior probabilities and the resultant classification posterior probabilities affect the models.

We used two published datasets that include multiple linear measurements for isolated theropod teeth and that were used as the basis for prior morphometric analyses (Hendrickx *et al*. 2015; Larson *et al*. 2016). These each include a wide range of theropod taxa, with broad spatial and temporal distributions. Each specimen has five measured morphometric variables that are simple 2D linear distances (or representations thereof) between repeatable landmarks on the tooth crowns (Fig. 2): crown base length (CBL), length of the base of the crown measured along its mesiodistal axis; crown base width (CBW), width of the base of the crown measured along its linguolabial axis perpendicular to the CBL; crown height (CH), height of the crown measured from the tip of the tooth to the base of the enamel; number of denticles per millimetre along the midpoint of the anterior carina (ADM); and number of denticles per millimetre along the midpoint of the posterior carina (PDM) (Currie *et al*. 1990; Smith *et al*. 2005; Larson and Currie 2013). These datasets comprise human-selected and hand measured morphometric data rather than measurements derived from photographic or other digital sources of information (such as CT-data) that have also been used in machine learning classifications (e.g. Hoyal Cuthill*, et al.* 2019). As such, it is inevitable that some degree of error will be introduced into the measurement process. However, given that the classification of isolated theropod teeth is a common requirement in vertebrate palaeontology, and the currently available datasets are all hand measured morphometric data, we feel there is value in applying such techniques to this data.

The Hendrickx *et al.* (2015) dataset consists of 995 individual cases belonging to 62 taxa from 19 major theropod clades (e.g., Megalosauridae, Tyrannosauridae, Dromaeosauridae, Abelisauridae) ranging in age from the Pliensbachian to the Maastrichtian with a global distribution. We analysed the data at two different taxonomic levels: a genus-level grouping of 680 cases and 32 classes and a higher-level clade aggregation comprising 886 cases and 14 classes. The dataset of Larson *et al.* (2016) comprises 3,104 maniraptoran theropod teeth from 18 lithostratigraphic units ranging in age from the uppermost Santonian (Milk River Formation) through to the Maastrichtian (Hell Creek Formation) of western North America. We analysed these data at two different taxonomic levels: a generic-level grouping containing 3020 cases and 17 classes; and a higher-level aggregation containing 3020 cases and four classes (Dromaeosauridae, Troodontidae, *Richardoestesia* and cf. Aves). We did not undertake a species level analysis due to the lack of species-level data with enough complete cases.

Data preparation

Prior to analysis we undertook a series of data exploration and general preparation steps. Each published dataset reports individual specimens at different taxonomic levels. For example, Hendrickx *et al.* (2015) list specimens at the generic level, whereas Larson *et al.* (2016) use species, with some of the latter split into stratigraphic units. To compare different models across both datasets, we aggregated groups of specimens to increasingly higher taxonomic levels. We removed any cases where it was unclear from the literature that a zero value in the data indicated a true zero (e.g., no anterior denticles) or represented missing data and, as some of the techniques applied require no missing data in the predictor variables, we removed all incomplete cases. Some classification techniques, such as LDA, are sensitive to the number of cases comprising individual groups in relation to the total number of predictor variables (Kuhn and Johnson 2013a; Zavorka and Perrett 2014) and require more cases per group than predictor variables. In addition, MacLeod (2018) noted that true group structures can be masked when the number of variables is greater than the number of cases. This is caused by having insufficient numbers of data points per group to describe the group structures correctly. At each taxonomic level tested, we therefore removed entire groups where the total number of group members was less than or equal to the number of predictor variables. As no dataset exhibited a multivariate normal distribution, the predictor variables were log-transformed with a constant value of one added prior to transformation to allow the log of true zero values.

For each taxonomic level tested we split the data into training and testing samples with a 80:20 ratio using the R package Caret (Kuhn 2008) which attempts to balance the class distributions within the training and testing sets. To optimise our models we undertook k-fold cross validation on the training set. Cross validation reduces the problems of underfitting, not capturing enough information in the model to accurately predict new data, and overfitting where the model performs well on the training set but does not generalise enough to perform well on new data (Hastie *et al.* 2009a). K-fold cross validation randomly divides the original data into k equally-sized subsamples. In this case we used a k-value of 10, so that the original training dataset is randomly divided into 10 subsamples. Nine subsamples are used as the training set and one as the testing set. This is then repeated 10 times such that each case forms part of a training set k-1 times and a testing set once. The model effectiveness is then averaged over each repeat to give a single overall model accuracy. We additionally ran the subsequent models on the retained testing samples, i.e., the samples not used to create the classification model, to provide more accurate assessments of the predictive accuracy of each model on unknown data.

Some of the models require specific parameters or preparation: for Naïve Bayes, in order to compensate for the non-independence of variables in our test data, we used PCA scores as input into the model rather than the original data; for random forests, our models used 2000 trees (to ensure model stability) and a range of m_{try} values from two to five; for C5.0, we ran models both with and without winnowing and set the model to stop the boosting process at 100 trials. We also generated a classifier ruleset for each model comprising simple if-then rules for the predicted class based on the input predictor variables.

All analyses were performed using R version 3.6.0 (R Core Team 2019) in R Studio (RStudio Team 2016) with the Caret package (Kuhn 2008) used for model generation. The following R packages were used for specific models or processes: UBL for synthetic data generation (Branco *et al.* 2016); missForest to introduce random missing data (Stekhoven and Buehlmann 2012; Stekhoven 2013); mice for data imputation (van Buuren and Groothuis-Oudshoorn 2011); MASS, C5.0 and randomForest for specific classification models (Liaw and Wiener 2002; Venables and Ripley 2002; Kuhn *et al.* 2018); and ggplot, gridextra, cowplot and ggalluvial for plotting functions (Wickham 2016; Auguie 2017; Brunson 2019; Wilke 2019).

Data balancing

A common issue with published datasets on tooth linear measurements is the unequal distribution of group members between distinct groups within the dataset. For example, the Larson *et al.* (2016) dataset contains 3020 specimens broken down into 17 generic groups. The distribution of group membership within these data ranges from 1176 individual cases to only six cases. As previously noted, groups defined by small numbers of cases suffer from the inability for the cases to correctly define the group structure. This imbalance also causes the performance of machine learning classifiers to be degraded as there is a bias towards the majority classes in an attempt to reduce the overall classification error. There are various methods that can be used to balance a dataset, all of which involve either the addition or removal of data points. Undersampling works on the majority classes, reducing the number of cases in each class in turn to create a more balanced dataset. This has the negative effect of removing informative data about these classes. Oversampling works on the minority classes by increasing the number of observations by replication. Whilst this does not result in information loss the implicit assumption is that the minority class structures are adequate to define those classes. We employed a methodology, Synthetic minority oversampling technique (SMOTE), that shifts the learning bias towards minority classes by generating synthetic data in these classes (Chawla *et al.* 2002). SMOTE oversamples the minority classes by creating new data points in feature-space randomly along a line joining an existing point to its nearest neighbours. We tested two scenarios to balance the training dataset to see if this resulted in a more accurate classification overall. First, random undersampling (i.e., removal) of the most populated classes combined with oversampling (by synthetic data generation) of the least populated classes to create a new dataset containing approximately the same number of overall cases as the original. Second, oversampling of the least populated classes to create an enlarged dataset with no undersampling of the most populated classes. We created these synthetic datasets based on the Larson *et al.* (2016) data at two different taxonomic levels running a number of different classifier models across the synthetic data to compare results to the original.

Dealing with missing data

Fossil datasets commonly contain incomplete morphometric information due to the nature of their preservation. Parts of a specimen may be missing due to breakage or wear, distortion as a result of geological processes may result in a measurement being suspect and therefore excluded, and the presence of host matrix can obscure particular features. The problem of missing data can be overcome either by deleting cases with missing values, using a variety of techniques to predict missing values based on the overall dataset, or by using a technique that is not reliant on complete cases. The first two techniques are problematic: deleting cases can remove useful information from the dataset, and replacing values with either mean substitution or values imputed from multiple regression has a tendency to distort the dataset and therefore the resultant classification (Schafer 1997; Feldesman 2002). Here we test different scenarios using the C5.0 tree-based classifier, which is not reliant on complete data. To look at the effects of missing data we used the Larson *et al.* (2016) dataset, which was edited to contain only complete cases. We then generated five new training datasets (Fig. 3) from this where we introduced increasing proportions of randomly generated missing data into the predictor variables (at 5, 10, 20, 30 and 50% levels) using the missForest package (Stekhoven and Buehlmann 2012; Stekhoven 2013). C5.0 classification models were then built for each of these new training datasets and applied to the retained testing data each time, allowing us to model changes in classification accuracy as the amount of missing data in the training set varied. We examined the effect of predicting missing values for each of the new training sets where we had previously introduced missing data using the MICE package (van Buuren and Groothuis-Oudshoorn 2011). For each training set containing missing data we created five imputed data sets that differ only in imputed missing values. We then built C5.0 classification models for each of these imputed datasets and stacked the results together to generate a training set containing imputed data. The imputed training set was fed into a secondary C5.0 model to provide the final classification (Fig. 3). Finally, we generated a C5.0 model using the original, complete Larson *et al.* (2016) dataset where we retained cases with missing data.

Prior and posterior probabilities

Bayesian classifiers use a prior probability distribution of group membership to calculate the posterior probability distribution, i.e., the resultant classification. The prior is essentially the probability that an observation comes from a particular group. There are three ways of defining prior probabilities: the prior probabilities are equal for all the groups, such that there is an equal chance that an observation can come from any group; the probabilities of group membership are proportional to the training dataset group observations; or the true group distribution is known (irrespective of the current dataset) and the priors can be defined explicitly to match this. The choice of prior will affect the outcome of classifications, especially when some group populations may be rare due to either unequal sampling or are a true reflection of the population under study (Zavorka and Perrett 2014). We modelled the effects of defining both equal and proportional priors on the final classification result.

Understanding how the final class assignment is made by a classifier is also important before any value can be attached to the result. Classifiers base their decisions on final class values on the calculated posterior probability for each class on a case-by-case basis. Classifiers that use ensemble techniques to arrive at a final result will still use posterior probability to assign classes within each of the models before creating the ensemble. The class assigned to a particular case is simply the class with the highest posterior probability. In some cases the results are fairly unequivocal, but in others a degree of caution is required. Take a simple example of a three class problem and two cases. Case one reports posterior probabilities of: Class A = 0.8, Class B= 0.1 and Class C = 0.1. Case two reports posterior probabilities of: Class A = 0.34, Class B = 0.33 and Class C = 0.33. Both cases are assigned to class A on the basis of the highest posterior probability, but it is clear from the results that the strength of the classification in case two is weak. Here we look at how the posterior probability varies on a case-by-case basis for a classification derived from an MDA model.

Ensemble classifier

For our ensemble classifier we combined the logistic regression, MDA and RF models as these employ differing techniques, with MDA and RF generally achieving the highest individual model accuracy (see Results, below). We used majority-voting and model stacking to combine the individual classification results and generate the final classification.

RESULTS

Comparison of classification models

Table 2 shows the overall accuracies of our models as applied to both the Hendrickx *et al.* (2015) and Larson *et al.* (2016) datasets. The top performers in each case are the non-linear MDA model and the decision tree based random forests and C5.0 models. Linear models (LDA and LR) perform poorly across both datasets as does the non-linear naïve Bayes model. Overall classification accuracy, irrespective of the model employed, increases as the number of classes decreases (Table 2). This increase in accuracy is as a result of true group structures being correctly described by having sufficient numbers of datapoints per group.

Using the Hendrickx *et al.* (2015) dataset we ran the classifiers at two taxonomic levels, the first a genus level with 32 classes and 680 cases and the second at a higher (family) taxonomic level with 14 classes and 886 cases. The 32-class model accuracies range from 59.2% (naïve Bayes) to 77.4% (MDA) accuracy. The 14-class models show an overall increase in classification accuracy with accuracies for the two highest performing models (random forests and C5.0) at around 80%. Compared to the equivalent Hendrickx *et al.* (2015) classification using LDA, the tree based models increase the overall accuracy of the prediction by around 10% from 70.9% as reported by Hendrickx *et al.* (2015) to 80.4% from the RF classifier. Our LDA model based on these data similarly boosts the overall accuracy to 76.7%. Figure 4 depicts the normalised confusion matrix for the Hendrickx *et al.* (2015) 14-class dataset from LDA, MDA, RF and C5.0 classifiers showing per-clade accuracies for each model. Two dimensional scatterplots of canonical variates obtained from MDA for the Hendrickx *et al.* (2015) dataset are shown in Figure 5A, which visually depict the group separations in discriminant space. The random forest classifier (Fig. 5B) demonstrates the decrease in error rates both overall and for most individual clades as the number of trees in the model increases. We ran all random forest models with 2,000 trees: however, the results indicate that little improvement in model performance is reached after 1,000 trees. The models used three randomly selected predictors (m_{try} value) for the 32-class dataset and two for the 14-class dataset. Figure 6 depicts the overall C5.0 model accuracies for the Hendrickx *et al.* (2015) dataset at a range of boosting iterations and using both winnowing and no winnowing. Across both taxonomic levels tested the overall accuracy settles down at around 25–30 boosting iterations. For the 32-class dataset the rules-based model using no winnowing improves the predictive accuracy slightly, for the 14-class dataset the rules-based model again shows a slight improvement in predictive accuracy irrespective of the use of winnowing.

Results from analysis of the Larson *et al*. (2016) dataset, again at two different taxonomic levels, broadly reinforce the previous analysis (Table 2). Decision trees and MDA return the highest classification accuracies with LDA performing relatively poorly. The difference between accuracies narrows as the number of groups in the data decreases and the numbers of cases making up each class increases. Accuracy for the 17-class dataset models ranges from 69.7% (LDA and NB) to 75% (RF) when applied to the testing data, with the 4-class dataset accuracies ranging from 93.3% (NB) to

96.3% (MDA). As with the previous dataset, the accuracy of classification increases as data is aggregated to higher and higher taxonomic levels. This increase in accuracy is reflective of the increasing certainty of the taxonomy, an increase in the number of cases making up the training groups and the removal of misclassification errors between closely related clades such as *Richardoestesia gilmorei* and *R. isosceles*, which have a tendency to classify to each other. Figures 7 and 8 depict the normalised confusion matrices for the 17- and 4-class Larson *et al*. (2016) datasets from the LDA, MDA, RF and C5.0 classifiers. Group separations in discriminant space obtained from the MDA classifier are shown in Figure 9A, the first two canonical variates are plotted that together account for around 93% of the total variation in each case. The random forest classifiers (Fig. 9B) again demonstrate the decrease in error rates as the number of trees in the model increases. The 4-class model overall accuracy and the accuracy of Troodontidae and Dromaeosauridae show little change after 250 trees but Aves and *Paronychodon* are unstable to around 1,000 trees. The 17-class model is noisier but again settles down at around 1,000 trees. Figure 10 depicts the overall C5.0 model accuracies and Figure 11 visualises one of the decision trees for the 4-class model. Across both taxonomic levels tested the tree-based model outperforms the rules based model although the difference between the two is minimal especially at the 4-class level. Winnowing of the predictor variables has a negative impact on the accuracy at 17 classes but little if any effect at the 4-class level. Boosting iterations settle at around 25 for the 4-class model and 50 for the 17-class model.

Data balancing

Figure 12A and Table 3 depict the changes in classification accuracy for LDA, MDA, RF and C5.0 models as we generated synthetic data in an attempt to balance the number of cases per class. The results show that attempting to balance class membership by either a combination of undersampling and oversampling (balanced results) or by oversampling alone produces significantly worse accuracy than no balancing.

Missing data

Table 4 summarises the results of introducing missing data at various percentage levels into the Larson *et al.* (2016) dataset and then using imputation to replace missing values. The classification accuracies decrease as the amount of missing data increases, with the 17-class model accuracy dropping off at a sharper rate than the 4-class model. The results indicate that the C5.0 classifier copes reasonably well with up to 20% missing data in some scenarios (Fig. 12B). The 4-class model accuracy decreases from 96.2% with no missing data to 93.9% with 20% missing data. Data imputation has a positive effect on the classification accuracies in the 4-class scenario with imputation at the 5% level slightly outperforming the original (no missing data) classifier. In the case of the 17-class models imputation has little effect on the classification accuracy with most imputed models showing a slightly lower accuracy rate than the models developed with missing data.

Prior and posterior probabilities

The effects of changing prior probabilities are summarised in Table 5 for LDA and MDA classifiers. Equal prior probabilities have the effect of increasing the bias towards smaller and potentially unstable groups reducing the overall accuracy of the model when compared to proportional priors. This is seen most markedly for the MDA classifier.

Posterior probabilities from the MDA classifier for 10 cases of the Larson *et al.* (2016) dataset are shown in Table 6. For most of the cases the classifier results in unambiguous predicted classes such as for cases 2–4 where the probability of the case classifying to Dromaeosauridae is 1.0. In other cases there is a degree of ambiguity as to the final class prediction. This is demonstrated by cases 1 and 8 where the final class prediction is only weakly supported (probabilities of 0.57 and 0.55, respectively).

Figure 13 shows the posterior probability mapping for the Larson *et al.* (2016) 17-class dataset. It is apparent from the overall map that clades such as *Richardoestesia* and *Troodon* have well supported final class prediction compared to *Acheroraptor* and *Bambiraptor*.

Ensemble classifiers

Table 7 summarises the accuracy achieved by stacking three different models to create an ensemble classifier and the accuracy of the majority vote ensemble. The stacking ensemble increases the overall classification accuracy in all cases with the exception of the Hendrickx *et al.* (2015) 32-class model. The increase in accuracy ranges from 0.5% for the Hendrickx *et al.* (2015) 14-class model to 1.1% for the Larson *et al.* (2016) 4-class data. The majority voting ensemble increases the overall model accuracy for the Larson *et al.* (2016) 4-class data to 97.5% (a similar level to the stacked ensemble) but is less successful for the other data analysed with either the individual classifiers or the stacked ensemble outperforming. Figure 14 shows how the classification of the Larson *et al.* (2016) dataset changes as a result of using different classifiers (LR, MDA, RF) and a majority vote ensemble classification based on all three individual classifiers. Clades such as *Pectinodon*, *Zapsalis*, *Paronychodon* and Aves have a relatively consistent classification outcome across all classifiers. This contrasts with many of the other dromaeosaurids which cross-classify depending on the chosen classification algorithm. Figure 14 also depicts an 'unknown' group in the final majority voting ensemble. This is where none of the constituent classifiers agreed on a final class and is an indication that there may be a sub-group present in the data that was incorrectly assigned a class in the training data.

DISCUSSION

Our results demonstrate that the non-linear and machine learning techniques we applied to hand-measured morphometric data derived from isolated theropod teeth consistently outperform LDA. When applying similar tests to anthropological data, Feldesman (2002) found that there was little difference between LDA and classification trees with LDA outperforming tree-based methods in some cases, whereas other authors (e.g., Holden *et al.* 2011; Finch *et al.* 2014) found LDA (and LR) to be the worst performers across a range of scenarios. This obviously raises the question of how to choose the most appropriate classifier to apply to a dataset. As pointed out by Feldesman (2002), unless the data meet all of the theoretical conditions of the technique in question then there must be a lack of confidence in the predictions delivered. At a minimum, therefore, we would stress the importance of applying more than one technique to test the classification. In most studies, decision trees such as RF and C5.0 have been shown to be among the best performers and have few (if any) prior assumptions regarding data structures. We therefore recommend that a decision tree approach (or MDA, another strong classifier) be either the primary classifier or at least used to test the classification returned from the chosen primary classifier. Ensemble classifiers can increase the predictive power over a single classifier and also offer the opportunity to reduce the risk of choosing the 'wrong' classifier and, where possible, we advocate their usage also (Dietterich 2001).

We also demonstrate that the choice of prior probability can affect the outcome of the classification. As the true population distributions of fossil taxa are unknown, and sampling of taxa is essentially opportunistic, a reasonable assumption is that the probability of a random observation coming from a particular group is equal across the groups under investigation. We accept that a choice of equal prior probabilities can increase the bias towards smaller and potentially unstable groups and reduce the overall accuracy of the model (Table 5). Nonetheless, we would recommend using equal priors, as with fossil taxa the true population is unknown and therefore the sample population cannot reflect reality. Rigorous data preparation to reduce the number of small unstable groups can help, but there is then a trade-off between overall model accuracy and the potential that a group may need to be excluded from the model. Datasets that contain missing data within the predictor variables complicate matters, as traditional LDA algorithms will not use incomplete cases. Our results indicate that imputing data as an alternative to deleting incomplete cases degrades the classifier accuracy substantially (Table 4; Fig. 12B). As decision trees can handle missing data we would recommend them over other alternatives as a first choice where the analysis of cases with missing data is a requirement. Class-imbalanced data biases the prediction towards majority groups and some techniques such as LDA perform badly with class imbalances. Our results suggest that using methods such as SMOTE to address this, by balancing class ratios via either synthetic case generation or under-sampling, degrades the classifier accuracy substantially (Fig. 12A). Blagus and Lusa (2013), however, concluded that whereas SMOTE was ineffective for discriminant analysis classifiers it may be of some benefit for other classifiers, such as decision trees. Although we would not rule out using synthetic data generation to balance classes, the effects of doing so need to be clearly understood (for example driving a bias towards the original minority classes) and the results tested against other classifiers using the imbalanced data. We would strongly recommend that posterior probabilities are checked as part of the process to verify the final classification.

Recent studies, such as Hendrickx *et al.* (2019), suggest that apomorphic character-based morphological data is potentially a more useful tool for distinguishing isolated theropod tooth crowns than morphometric data. However, we show that the careful application of machine learning techniques using the frameworks discussed in this study demonstrate that continuous quantitative morphometric data can also discriminate isolated theropod teeth with taxonomic accuracy of up to 96% in the specific datasets we used. The use of appropriate multiple classifiers coupled with a considered approach and understanding of the effects of missing data, initial group sizes and class imbalances are an improvement on the current commonly used techniques and yield rapid and statistically robust group predictions. Classification of isolated teeth in this manner will improve with better data, namely more cases per clade, to train the classifiers on. The careful addition of new measurement variables may also improve classification accuracies. As machine learning techniques have already been shown to be able to successfully classify taxa even with evolutionary convergence (e.g., Hoyal Cuthill *et al*., 2019) it is likely that even highly heterodont theropod clades and clades exhibiting dental morphological convergence could be accurately distinguished given the right amount of data and careful pre-processing of the data. It is probable that in some circumstances a combination of a dentition-based cladistic analysis and morphometric analysis may achieve the best results. The taxon-level grouping that is chosen will have an impact on the overall accuracy of the model simply because this controls the number of cases per group which in turn impacts on the ability of the classifier to accurately describe that group. An attempt to classify at a species level where each species is described by, for example, four individual teeth will be less accurate than a genus level classification where each genus is represented by several hundred teeth.

CONCLUSIONS

In order to assess the performance of machine learning techniques on basic morphometric data derived from isolated theropod dinosaur tooth crowns a comparative study was undertaken using two published datasets. Various machine learning procedures were applied to each dataset in order to test the predictive accuracy under a range of different conditions. The results presented here, although specific to the tested datasets, demonstrate several important points:

1. Although LDA was generally the poorest performer in terms of accuracy, its predictive capability improved with larger class sizes.
2. Data subjected to predictive classification techniques needs to be rigorously assessed prior to classification for normality, missing data, class imbalances and class size. If data fail these tests then alternatives to LDA need to be considered.
3. Decision tree techniques such as random forest and C5.0 consistently outperformed other methods and we would advocate their usage for such classification problems.
4. Attempts to balance classes either by synthetic data generation, or by over- or undersampling of classes, significantly degraded the classification accuracy and care must be taken before employing these techniques.
5. Increasing percentages of missing data and the use of imputation to correct for this caused steep decreases in the predictive accuracy of those classifiers designed to handle such data (e.g., C5.0).
6. Different classifiers will assign the same case to different classes. The use of ensemble classifiers and an assessment of the resultant posterior probabilities helps to reduce the possibility of the ‘wrong’ technique being chosen.

As a result of this study we would recommend the use of decision trees as an alternative approach to LDA. The final aim of the analysis should guide the choice of random forest or C5.0. If the goal is to predict the taxon that a tooth falls into then random forests are a good choice. If the aim is to classify and to be able to see how the classification is built within the tree structure then C5.0 should be used. In practice we would recommend corroboration of any results by checking predictions with another technique, preferably via the use of ensemble classifiers. The use of such techniques on isolated theropod teeth demonstrates that high levels of predictive taxonomic accuracy are possible from simple morphometric data as long as care is taken to understand the structure of the data in question and the assumptions that various techniques require.

Acknowledgements. We are grateful to Natalie Cooper whose comments, suggestions and corrections on an earlier draft helped to improve this manuscript. Norman MacLeod is thanked for fruitful early discussion. We would also like to thank Eva Larsson for useful comments on early versions of the manuscript. We thank the two reviewers, C. Hendrickx and J. Hoyal Cuthill, whose comments and suggestions helped to improve and clarify this manuscript.

REFERENCES

AFANADOR, N. L., SMOLINSKA, A., TRAN, T. N. and BLANCHET, L. 2016. Unsupervised random forest: a tutorial with case studies. *Journal of Chemometrics*, **30**, 232–241.

ARGAST, S., FARLOW, J. O., GABET, R. M. and BRINKMAN, D. L. 1987. Transport-induced abrasion of fossil reptilian teeth: Implications for the existence of Tertiary dinosaurs in the Hell Creek Formation, Montana. *Geology*, **15**, 927–930.

AUGUIE, B. 2017. gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3.

BARRETT, P. M., BUTLER, R. J., MUNDIL, R., SCHEYER, T. M., IRMIS, R. B. and SÁNCHEZ-VILLAGRA, M. R. 2014. A palaeoequatorial ornithischian and new constraints on early dinosaur diversification. *Proceedings of the Royal Society of London B: Biological Sciences*, **281**, 20141147.

BASZIO, S. 1997. Systematic palaeontology of isolated dinosaur tooth from the latest Cretaceous of South Alberta, Canada. *Courier Forschungsinstitut Senckenberg*, **196**, 33–77.

BECERRA, M. G., POL, D., MARSICANO, C. A. and RAUHUT, O. W. M. 2013. The dentition of Manidens condorensis (Ornithischia; Heterodontosauridae) from the Jurassic Cañadón Asfalto Formation of Patagonia: morphology, heterodonty and the use of statistical methods for identifying isolated teeth. *Historical Biology*, **26**, 480–492.

BLAGUS, R. and LUSA, L. 2013. SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics*, **14**, 1–16.

BOYD, C. A. 2015. The systematic relationships and biogeographic history of ornithischian dinosaurs. *PeerJ*, **3**, e1523.

BRANCO, P., RIBEIRO, R. P. and TORGO, L. 2016. UBL: an R package for utility-based learning. *CoRR*, **abs/1604.08079**.

BREIMAN, L. 2001. Random Forests. *Machine Learning*, **45**, 5–32.

BREIMAN, L., FRIEDMAN, J., STONE, C. J. and OLSHEN, R. A. 1984. *Classification and regression trees*. Chapman and Hall / CRC Press, London, 368 pp.

BRITZKE, E. R., DUCHAMP, J. E., MURRAY, K. L., SWIHART, R. K. and ROBBINS, L. W. 2011. Acoustic identification of bats in the eastern United States: A comparison of parametric and nonparametric methods. *The Journal of Wildlife Management*, **75**, 660–667.

BRUNSON, J. C. 2019. ggalluvial: Alluvial Diagrams in 'ggplot2'. R package version 0.9.1.

CHANDLER, L. 1990. Taxonomic and functional significance of serrated tooth morphology in theropod dinosaurs. Unpublished MS thesis, Yale University, New Haven, Connecticut, 163 pp.

CHARIG, A. J. and CROMPTON, A. W. 1974. The alleged synonymy of *Lycorhinus* and *Heterodontosaurus*. *Annals of the South African Museum*, **64**, 167–189.

CHAWLA, N. V., BOWYER, K. W., HALL, L. O. and KEGELMEYER, W. P. 2002. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, **16**, 321–357.

CHEN, X. and ISHWARAN, H. 2012. Random forests for genomic data analysis. *Genomics*, **99**, 323–329.

CHRISTIN, S., HERVET, É. and LECOMTE, N. 2019. Applications for deep learning in ecology. *Methods in Ecology and Evolution*, **10**, 1632–1644.

COOMBS, W. P. 1990. Teeth and taxonomy in ankylosaurs. *In* CARPENTER, K. and CURRIE, P. J. (eds). *Dinosaur systematics: approaches and perspectives*. Cambridge University Press, Cambridge, 318 pp.

CORENTIN, B. and SALVADOR, B. 2018. A New Fossil Species of *Boa* Linnaeus, 1758 (Squamata, Boidae), from the Pleistocene of Marie-Galante Island (French West Indies). *Journal of Vertebrate Paleontology*, **38:3**. doi: 10.1080/02724634.2018.1462829

COURONNÉ, R., PROBST, P. and BOULESTEIX, A.-L. 2018. Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinformatics*, **19**, 1–14.

CRIMINISI, A., SHOTTON, J. and KONUKOGLU, E. 2012. Decision forests: a unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and trends in computer graphics and vision*, **7**, 81–227.

CURRIE, P. J., RIGBY, J. K. and SLOAN, R. E. 1990. Theropod teeth from the Judith River Formation of Southern Alberta, Canada. *In* CARPENTER, K. and CURRIE, P. J. (eds). *Dinosaur Systematics Approaches and Perspectives*. Cambridge University Press, Cambridge, 318 pp.

CUTLER, D. R., EDWARDS JR, T. C., BEARD, K. H., CUTLER, A., HESS, K. T., GIBSON, J. and LAWLER, J. J. 2007. Random forests for classification in ecology. *Ecology*, **88**, 2783–2792.

DE SPIEGELEER, J., MADAN, D. B., REYNERS, S. and SCHOUTENS, W. 2018. Machine learning for quantitative finance: fast derivative pricing, hedging and fitting. *Quantitative Finance*, **18**, 1635–1643.

DIETTERICH, T. G. 2001. Ensemble methods in machine learning. *In* KITTLER, J. and ROLI, F. (eds). *Multiple classier systems*. Springer-Verlag, Berlin, 404 pp.

DODSON, P. and DAWSON, S. 1991. Making the fossil record of dinosaurs. *Modern Geology*, **16**, 3–15.

DŽEROSKI, S. 2001. Applications of symbolic machine learning to ecological modelling. *Ecological Modelling*, **146**, 263–273.

EVANS, A. R., WILSON, G. P., FORTELIUS, M. and JERNVALL, J. 2007. High-level similarity of dentitions in carnivorans and rodents. *Nature*, **445**, 78–81.

EVANS, S. E. and MILNER, A. R. 1994. Middle Jurassic microvertebrate assemblages from the British Isles. *In* FRASER, N. C. and SUES, H.-D. (eds). *In the shadow of the dinosaurs. Early Mesozoic tetrapods*. Cambridge University Press, Cambridge, 435 pp.

FANTI, F. and THERRIEN, F. 2007. Theropod tooth assemblages from the Late Cretaceous Maevarano Formation and the possible presence of dromaeosaurids in Madagascar. *Acta Palaeontologica Polonica*, **52**, 155–166.

FARLOW, J. O., BRINKMAN, D. B., ABLER, W. L. and CURRIE, P. J. 1991. Size, shape and serration density of theropod dinosaur lateral teeth. *Modern Geology*, **16**, 161–198.

FELDESMAN, M. R. 2002. Classification trees as an alternative to linear discriminant analysis. *American Journal of Physical Anthropology*, **119**, 257–275.

FINCH, W. H., BOLIN, J. H. and KELLEY, K. 2014. Group membership prediction when known groups consist of unknown subgroups: a Monte Carlo comparison of methods. *Frontiers in Psychology*, **5**, 1–12.

FIORILLO, A. R. and CURRIE, P. J. 1994. Theropod Teeth from the Judith River Formation (Upper Cretaceous) of South-Central Montana. *Journal of Vertebrate Paleontology*, **14**, 74–80.

FISHER, R. A. 1936. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**, 179–188.

FRASER, D. and THEODOR, J. M. 2011. Anterior dentary shape as an indicator of diet in ruminant artiodactyls. *Journal of Vertebrate Paleontology*, **31**, 1366–1375.

GATES, T. A., ZANNO, L. E. and MAKOVICKY, P. J. 2015. Theropod teeth from the upper Maastrichtian Hell Creek Formation "Sue" Quarry: New morphotypes and faunal comparisons. *Acta Palaeontologica Polonica*, **60**, 131–139.

GERKE, O. and WINGS, O. 2016. Multivariate and Cladistic Analyses of Isolated Teeth Reveal Sympatry of Theropod Dinosaurs in the Late Jurassic of Northern Germany. *PLOS ONE*, **11**, e0158334.

HASTIE, T. and TIBSHIRANI, R. 1996. Discriminant Analysis by Gaussian Mixtures. *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**, 155–176.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. 2009a. Springer series in statistics. *The elements of statistical learning*. Springer-Verlag, New York, 745 pp.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. 2009b. Ensemble learning. In HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (eds). The elements of statistical learning. Springer-Verlag, New York, 745 pp.

HECKERT, A. B. 2002. A revision of the Upper Triassic ornithischain dinosaur *Revueltosaurus,* with a description of a new species. *New Mexico Museum of Natural History and Science*, **21**, 253–267.

--- 2004. Late triassic microvertebrates from the Lower Chinle Group (Otischalkian-Adamanian:Carnian), southwestern USA. *New Mexico Museum of Natural History and Science*, **27**, 1–170.

HENDRICKX, C. and MATEUS, O. 2014. Abelisauridae (Dinosauria: Theropoda) from the Late Jurassic of Portugal and dentition-based phylogeny as a contribution for the identification of isolated theropod teeth. *Zootaxa*, **3759**, 1–74.

HENDRICKX, C., MATEUS, O. and ARAÚJO, R. 2015. The dentition of megalosaurid theropods. *Acta Palaeontologica Polonica*, **60**, 627–642.

HENDRICKX, C., MATEUS, O., ARAÚJO, R. and CHOINIERE, J. 2019. The distribution of dental features in non-avian theropod dinosaurs: Taxonomic potential, degree of homoplasy, and major evolutionary trends. *Palaeontologia Electronica*, **22.3.74**, 1–110

HENDRICKX, C., TSCHOPP, E. and EZCURRA, M. D. 2020. Taxonomic identification of isolated theropod teeth: The case of the shed tooth crown associated with Aerosteon (Theropoda: Megaraptora) and the dentition of Abelisauridae. Cretaceous Research, 108, 104312. doi: 10.1016/j.cretres.2019.104312

HOFFMAN, D. K., EDWARDS, H. R., BARRETT, P. M. and NESBITT, S. J. 2019. Reconstructing the archosaur radiation using a Middle Triassic archosauriform tooth assemblage from Tanzania. *PeerJ*, **7**, e7970.

HOLDEN, J. E., FINCH, W. H. and KELLEY, K. 2011. A Comparison of Two-Group Classification Methods. *Educational and Psychological Measurement*, **71**, 870–901.

HOYAL CUTHILL, J. F., GUTTENBERG, N., LEDGER, S., CROWTHER, R. and HUERTAS, B. 2019. Deep learning on butterfly phenotypes tests evolution's oldest mathematical model. Science Advances, 5, eaaw4967. doi: 10.1126/sciadv.aaw4967

HUNT, A. and LUCAS, S. 1994. Ornithischian dinosaurs from the Upper Triassic of the United States. *In* FRASER, N. C. and SUES, H.-D. (eds). *In the shadow of the dinosaurs. Early Mesozoic tetrapods*. Cambridge University Press, Cambridge, 435 pp.

JOLLIFFE, I. T. 2002. *Principle Component Analysis*. Springer, New York.

KER, J., WANG, L., RAO, J. and LIM, T. 2018. Deep learning applications in medical image analysis. *IEEE Access*, **6**, 9375–9389.

KNOLL, F. and RUIZ-OMEÑACA, J. I. 2009. Theropod teeth from the basalmost Cretaceous of Anoual (Morocco) and their palaeobiogeographical significance. *Geological Magazine*, **146**, 602–616.

KUHN, M. 2008. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, **1**, 1–26.

KUHN, M. and JOHNSON, K. 2013a. *Applied predictive modelling*. Springer-Verlag, New York, 600 pp.

KUHN, M. and JOHNSON, K. 2013b. Classification trees and rule-based models. *In* KUHN, M. and JOHNSON, K. (eds). *Applied predictive modelling*. Springer-Verlag, New York, 600 pp.

KUHN, M. and JOHNSON, K. 2013c. Measuring performance in classificaiton models. *In* KUHN, M. and JOHNSON, K. (eds). *Applied predictive modelling*. Springer-Verlag, New York, 600 pp.

KUHN, M., WESTON, S., CULP, M., COULTER, N. and QUINLAN, R. 2018. *C5.0 decision trees and rule-based model*. RuleQuest Research Pty Ltd.

LARSON, D. W. 2008. Diversity and variation of theropod dinosaur teeth from the uppermost Santonian Milk River Formation (Upper Cretaceous), Alberta: a quantitative method supporting identification of the oldest dinosaur tooth assemblage in Canada. *Canadian Journal of Earth Sciences*, **45**, 1455–1468.

LARSON, DEREK W., BROWN, CALEB M. and EVANS, DAVID C. 2016. Dental Disparity and Ecological Stability in Bird-like Dinosaurs prior to the End-Cretaceous Mass Extinction. *Current Biology*, **26**, 1325–1333.

LARSON, D. W. and CURRIE, P. J. 2013. Multivariate Analyses of Small Theropod Dinosaur Teeth and Implications for Paleoecological Turnover through Time. *PLoS ONE*, **8**, e54329.

LIAW, A. and WIENER, M. 2002. Classification and Regression by randomForest. *R News*, **2**, 18–22.

MA, X. and LV, S. 2019. Financial credit risk prediction in internet finance driven by machine learning. *Neural Computing and Applications*, **31**, 8359–8367.

MACLEOD, N. 2007. The Systematics Association Special Volume Series 74. *Automated taxon identification in systematics : theory, approaches and applications*. CRC Press, New York.

--- 2015. The direct analysis of digital images (eigenimage) with a comment on the use of discriminant analysis in morphometrics. 156–182. *In* LESTREL, P. E. (ed.) *Proceedings of the Third International Symposium on Biological Shape Analysis*. World Scientific, Singapore,

--- 2017. On the Use of Machine Learning Methods in Morphometric Analysis. 134–171. *In* LESTREL, P. E. (ed.) *Biological Shape Analysis, Proceedings of the 4th International Symposium on Biological Shape Analysis (ISBSA)*. World Scientific, Singapore, School of Dentistry, UCLA, USA, 19 – 22 June 2015.

--- 2018. The quantitative assessment of archaeological artifact groups: Beyond geometric morphometrics. *Quaternary Science Reviews*, **201**, 319–348.

--- 2019. Artificial intelligence and machine learning in the earth sciences. *Acta Geologica Sinica (English Edition)*, **93**, 48–51.

MARRAMÀ, G. and KRIWET, J. 2017. 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*, **12**, e0188806.

MARSLAND, S. 2015. *Machine learning: An algorithmic perspective*. CRC Press, Boca Raton, Florida

MAUGIS, C., CELEUX, G. and MARTIN-MAGNIETTE, M. L. 2011. Variable selection in model-based discriminant analysis. *Journal of Multivariate Analysis*, **102**, 1374–1387.

MEHTA, S. and SHUKLA, D. 2015. Optimization of C5.0 classifier using Bayesian theory. 1–6. *2015 International Conference on Computer, Communication and Control (IC4)*.

MELSTROM, K. M. and IRMIS, R. B. 2019. Repeated Evolution of Herbivorous Crocodyliforms during the Age of Dinosaurs. *Current Biology*, **29**, 2389–2395.e3. doi: 10.1016/j.cub.2019.05.076

METCALF, S. J. and WALKER, R. J. 1994. A new bathonian microvertebrate locality in the English Midlands. *In* FRASER, N. C. and SUES, H.-D. (eds). *In the shadow of the dinosaurs. Early Mesozoic tetrapods*. Cambridge University Press, Cambridge, 435 pp.

MILLA CARMONA, P. S., LAZO, D. G. and SOTO, I. M. 2016. Giving taxonomic significance to morphological variability in the bivalve Ptychomya Agassiz. *Palaeontology*, **59**, 139–154.

MONSON, T. A., ARMITAGE, D. W. and HLUSKO, L. J. 2018. Using machine learning to classify extant apes and interpret the dental morphology of the chimpanzee-human last common ancestor. *PaleoBios*, **35**, 1–20.

ONOJEGHUO, A. O., BLACKBURN, G. A., WANG, Q., ATKINSON, P. M., KINDRED, D. and MIAO, Y. 2018. Mapping paddy rice fields by applying machine learning algorithms to multi-temporal Sentinel-1A and Landsat data. *International Journal of Remote Sensing*, **39**, 1042–1067.

ŐSI, A., PRONDVAI, E., MALLON, J. and BODOR, E. R. 2016. Diversity and convergences in the evolution of feeding adaptations in ankylosaurs (Dinosauria: Ornithischia). *Historical Biology*, **29**, 539–570.

OSTROM, J. H. and WELLNHOFER, P. 1990. *Tricerotops*: an example of flawed systematics. *In* CARPENTER, K. and CURRIE, P. J. (eds). *Dinosaur Systematics Approaches and Perspectives*. Cambridge University Press, Cambridge, 318 pp.

PADIAN, K. 1990. The Ornithischian Form Genus Revueltosaurus from the Petrified Forest of Arizona (Late Triassic: Norian; Chinle Formation). *Journal of Vertebrate Paleontology*, **10**, 268–269.

QUINLAN, J. R. 1993. *C4.5: Programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 301 pp.

R CORE TEAM 2019. R: A Language and Environment for Statistical Computing.

RAUHUT, O. W. M. 2002. Dinosaur teeth from the Barremian of Uña, Province of Cuenca, Spain. *Cretaceous Research*, **23**, 255–263.

RAUSCH, J. R. and KELLEY, K. 2009. A comparison of linear and mixture models for discriminant analysis under nonnormality. *Behavior Research Methods*, **41**, 85–98.

RIFFENBURGH, R. H. 2012. Chapter 19 - Modeling Concepts and Methods. *In* RIFFENBURGH, R. H. (ed.) *Statistics in Medicine (Third Edition)*. Academic Press, San Diego, 690 pp.

ROLI, F., GIACINTO, G. and VERNAZZA, G. 2001. Multiple Classifier Systems. *Methods for Designing Multiple Classifier Systems*. Springer, Berlin, Heidelberg, 456 pp.

RSTUDIO TEAM 2016. RStudio: Integrated Development Environment for R.

RUSSELL, S. and NORVIG, P. 2009. *Artificial intelligence: A modern approach*. Prentice Hall Press, New Jersey, 1152 pp.

SALZBERG, S. L. 1994. C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. *Machine Learning*, **16**, 235-240.

SAMMAN, T., POWELL, G. L., CURRIE, P. J. and HILLS, L. V. 2005. Morphometry of the teeth of western North American tyrannosaurids and its applicability to quantitative classification. *Acta Palaeontol. Pol.*, **50**, 757–776.

SANKEY, J. T., BRINKMAN, D., GUENTHER, M. and CURRIE, P. J. 2002. Small theropod and bird teeth from the late Cretaceous (late Campanian) Judith River Group, Alberta. *Journal of Paleontology*, **76**, 751–763.

SCHAFER, J. L. 1997. *Analysis of incomplete multivariate data*. Chapman & Hall, New York, 444 pp.

SCHUETTPELZ, E., FRANDSEN, P. B., DIKOW, R. B., BROWN, A., ORLI, S., PETERS, M., METALLO, A., FUNK, V. A. and DORR, L. J. 2017. Applications of deep convolutional neural networks to digitized natural history collections. *Biodiversity Data Journal*, **5**, e21139.

SERENO, P. C. 1991. Lesothosaurus, “Fabrosaurids,” and the early evolution of Ornithischia. *Journal of Vertebrate Paleontology*, **11**, 168–197.

SHI, T. and HORVATH, S. 2006. Unsupervised learning with random forest predictors. *Journal of Computational and Graphical Statistics*, **15**, 118–138.

SMITH, J. B. 2005. Heterodonty in Tyrannosaurus rex: Implications for the taxanomic and systematic utility of Theropod dentitions. *Journal of Vertebrate Paleontology*, **25**, 865–887.

SMITH, J. B., VANN, D. R. and DODSON, P. 2005. Dental morphology and variation in Theropod Dinosaurs: Implications for the taxanomic identification of isolated teeth. *The Anatomical Record Part A*, **285A**, 699–736.

SON, N.-T., CHEN, C.-F., CHEN, C.-R. and MINH, V.-Q. 2018. Assessment of Sentinel-1A data for rice crop classification using random forests and support vector machines. *Geocarto International*, **33**, 587–601.

STEKHOVEN, D. J. 2013. missForest: Nonparametric Missing Value Imputation using Random Forest.

STEKHOVEN, D. J. and BUEHLMANN, P. 2012. MissForest - non-parametric missing value imputation for mixed-type data. *Bioinformatics*, **28**, 112–118.

STRICKSON, E., PRIETO-MÁRQUEZ, A., BENTON, M. J. and STUBBS, T. L. 2016. Dynamics of dental evolution in ornithopod dinosaurs. *Scientific Reports*, **6**, (28904).

THULBORN, R. A. 1973. Teeth of ornithischian dinosaurs from the Upper Jurassic of Portugal, with description of a hypsilophodontid (*Phyllodon henkeli* gen. et sp. nov.) from the Guimarota lignite. *Contribuiçao para o conhecimento da Fauna do Kimeridgiano da Mina de Lignito Guimarota (Leiria, Portugal) III Parte, Serviços Geológicos de Portugal, Memória 22 (Nova Série)*. 469 pp.

--- 1992. Taxonomic characters of *Fabrosaurus australis,* an ornithischian dinosaur from the Lower Jurassic of Southern Africa. *Geobios*, **25**, 283–292.

TORICES, A., REICHEL, M. and CURRIE, P. J. 2014. Multivariate analysis of isolated tyrannosaurid teeth from the Danek Bonebed, Horseshoe Canyon Formation, Alberta, Canada. *Canadian Journal of Earth Sciences*, **51**, 1045–1051.

VALIANT, L. 1984. A theory of the learnable. *Communications of the ACM*, **27**, 1134-1142.

VAN BUUREN, S. and GROOTHUIS-OUDSHOORN, K. 2011. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, **45**, 1–67.

VENABLES, W. N. and RIPLEY, B. D. 2002. *Modern Applied Statistics with S*. Springer, New York.

WELCH, B. L. 1939. Note on discriminant functions. *Biometrika*, **31**, 218–220.

WHITENACK, L. B. and GOTTFRIED, M. D. 2010. A Morphometric Approach for Addressing Tooth-Based Species Delimitation in Fossil Mako Sharks, Isurus (Elasmobranchii: Lamniformes). *Journal of Vertebrate Paleontology*, **30**, 17–25.

WICKHAM, H. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York, 260 pp.

WILKE, C. O. 2019. cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'. R package version 0.9.4.

WILLIAMSON, T. E. and BRUSATTE, S. L. 2014. Small Theropod Teeth from the Late Cretaceous of the San Juan Basin, Northwestern New Mexico and Their Implications for Understanding Latest Cretaceous Dinosaur Evolution. *PLoS ONE*, **9**, e93190.

WILSON, G. P., EVANS, A. R., CORFE, I. J., SMITS, P. D., FORTELIUS, M. and JERNVALL, J. 2012. Adaptive radiation of multituberculate mammals before the extinction of dinosaurs. *Nature*, **483**, 457–460.

WINGS, O., TÜTKEN, T., FOWLER, D., MARTIN, T., PFRETZSCHNER, H.-U. and SUN, G. 2015. Dinosaur teeth from the Jurassic Qigu and Shishugou Formations of the Junggar Basin (Xinjiang/China) and their paleoecologic implications. *Paläontologische Zeitschrift*, **89**, 485–502.

YOUNG, C. M. E., HENDRICKX, C., CHALLANDS, T. J., FOFFA, D., ROSS, D. A., BUTLER, I. B. and BRUSATTE, S. L. 2019. New theropod dinosaur teeth from the Middle Jurassic of the Isle of Skye, Scotland. *Scottish Journal of Geology*, **55**, 7–19.

ZAVORKA, S. and PERRETT, J. J. 2014. Minimum sample size considerations for two-group linear and quadratic discriminant analysis with rare populations. *Communications in Statistics - Simulation and Computation*, **43**, 1726–1739.

TABLE CAPTIONS

TABLE 1. Glossary of terms used in machine learning and classification.

TABLE 2. Classification results for different models using the Hendrickx *et al.* (2015) and Larson *et al.* (2016) datasets. Accuracies are shown for both the classification model and the testing data. LDA, linear discriminant analysis; LR, logistic regression; MDA, mixture discriminant analysis; NB, naïve Bayes; RF, random forests; C5.0, rule-based decision tree.

TABLE 3. Classification accuracy results for synthetic data generation (SMOTE) compared to unbalanced data for LDA, MDA, RF and C5.0 classifiers. Accuracy based on Larson *et al*. (2016) data.

TABLE 4. C5.0 classifier results on missing and imputed data for Larson *et al*. (2016) dataset.

TABLE 5. Effect of different prior probabilities on model accuracy.

TABLE 6. Posterior probabilities for 10 cases selected at random from the MDA classifier using the Larson *et al*. (2016) 4-class dataset.

TABLE 7. Ensemble model accuracy using model stacking. Accuracies are shown for both the individual models that make up the ensembles and the stacked and majority vote ensembles.

FIGURE CAPTIONS

FIG. 1. Hypothetical decision tree for the example of catching a train to London Victoria station.

FIG. 2. Tooth measurements used in this study. ADM, anterior denticles per millimetre; CBL, crown base length; CBW, crown base width; CH, crown height; PDM, posterior denticles per millimetre.

FIG. 3. Workflow for looking at the effect of missing data on predictive accuracy. A. Generating new datasets with missing data inserted at random. For this exercise we added missing data into the predictor variables at 5, 10, 20, 30 and 50% levels. B. Replacing missing data with substituted values. For the sake of clarity we have only shown the workflow for one of the training datasets containing missing data. This dataset was derived from workflow A.

FIG. 4. Normalised confusion matrices for LDA, MDA, RF and C5.0 classification models based on the Hendrickx *et al.* (2015) 14-class dataset. Reference classes are plotted on the x-axis and predicted classes on the y-axis.

FIG. 5. Hendrickx *et al.* (2015) 14-class dataset A. MDA canonical variates showing group separations in discriminant space. B. Random forest error rate per taxon and overall (OOB) classification error rate. For the sake of clarity only five taxa are shown on the RF plot.

FIG. 6. C5.0 accuracy plots for Hendrickx *et al.* (2015) data showing the effects of winnowing predictor variables and the rules vs. tree based models at different boosting iterations. A. 32-class model. B. 14-class model.

FIG. 7. Normalised confusion matrices for LDA, MDA, RF and C5.0 classification models based on the Larson *et al.* (2016) 17-class dataset. Reference classes are plotted on the x-axis and predicted classes on the y-axis. Pmx: pre-maxillary tooth.

FIG. 8. Normalised confusion matrices for LDA, MDA, RF and C5.0 classification models based on the Larson *et al.* (2016) 4-class dataset. Reference classes are plotted on the x-axis and predicted classes on the y-axis.

FIG. 9. A. MDA canonical variates plots for Larson *et al.* (2016) data showing group separations in discriminant space. B. Random forest error rate for Larson *et al.* (2016) 4-class model. Pmx: pre-maxillary tooth.

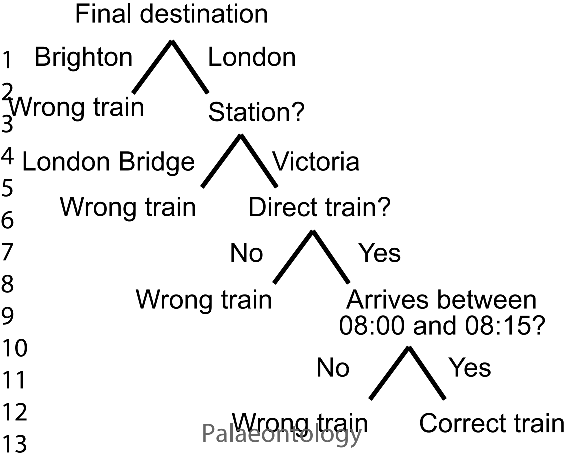
FIG. 10. C5.0 models for Larson *et al.* (2016) data showing the effects of winnowing predictor variables and the rules vs. tree based models at different boosting iterations. A. 17-class model. B. 4-class model.

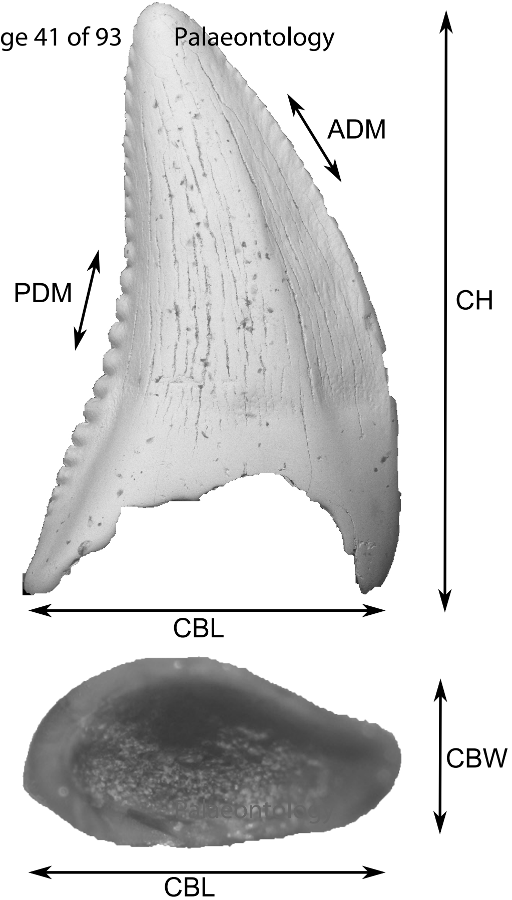
FIG. 11. Extract from the decision tree classifier Larson *et al.* (2016) data. Each node shows: the predicted class; the predicted probability of each class; the percentage of observations in each node.

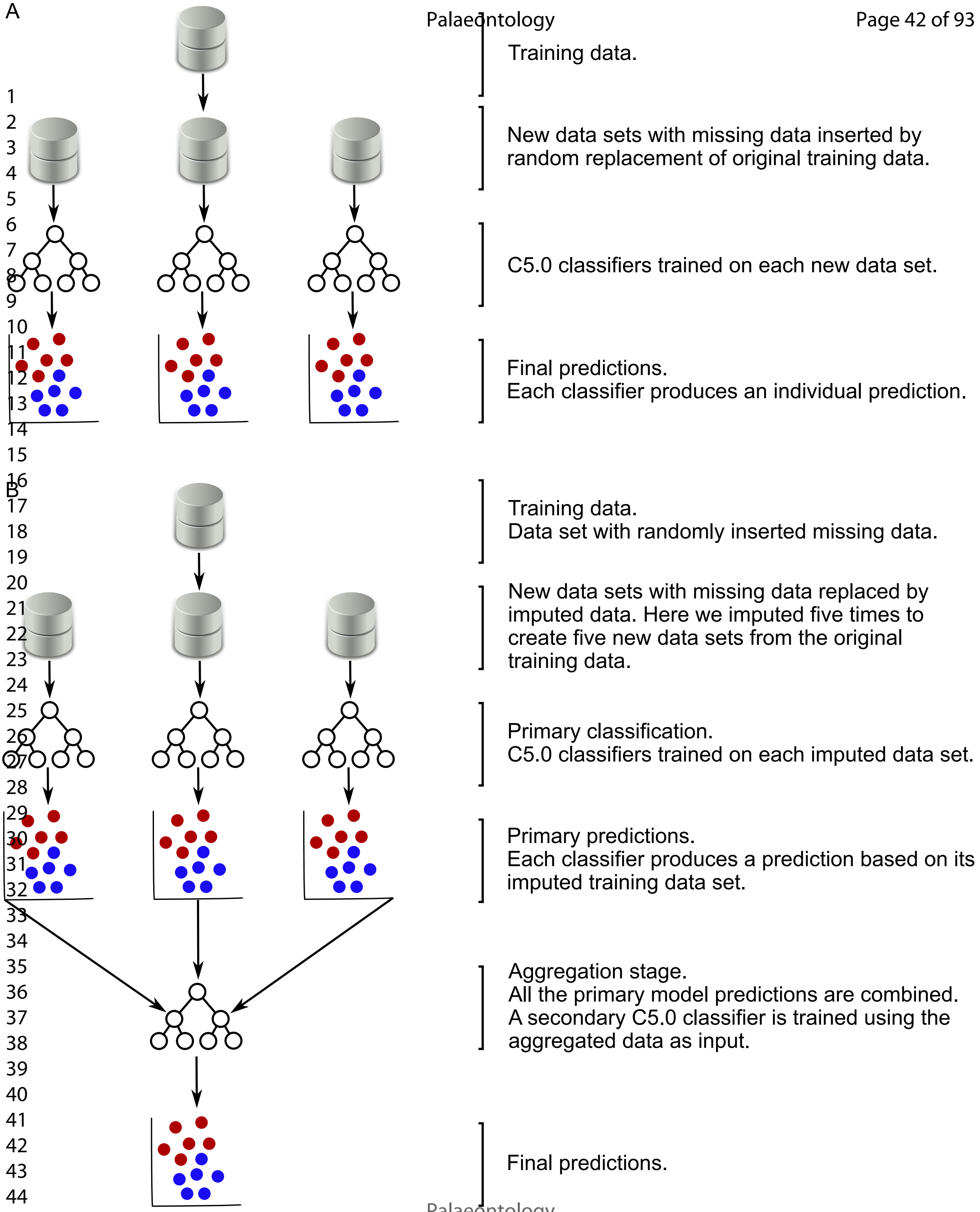
FIG. 12. A. C5.0 classifier accuracy for synthetically generated class balanced datasets B. C5.0 classifier accuracies for missing and imputed data at different levels. Horizontal dotted lines show the C5.0 model accuracy with no missing or imputed data.

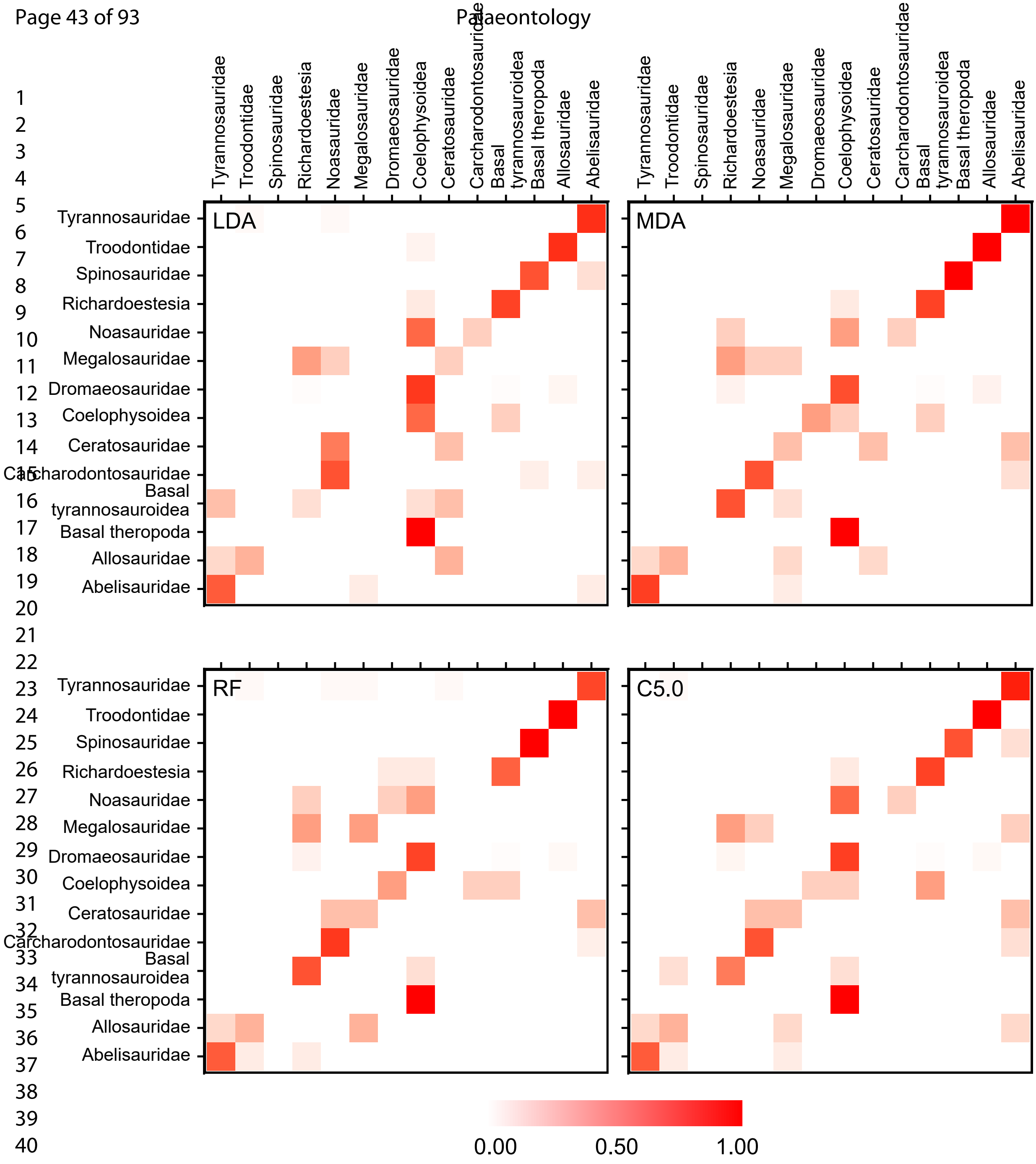
FIG. 13. Posterior probability heatmap. MDA classifier, Larson *et al.* (2016) 17-class dataset. A. Entire test dataset. B. First 30 cases. Each block on the x-axis represents one case. Pmx: pre-maxillary tooth.

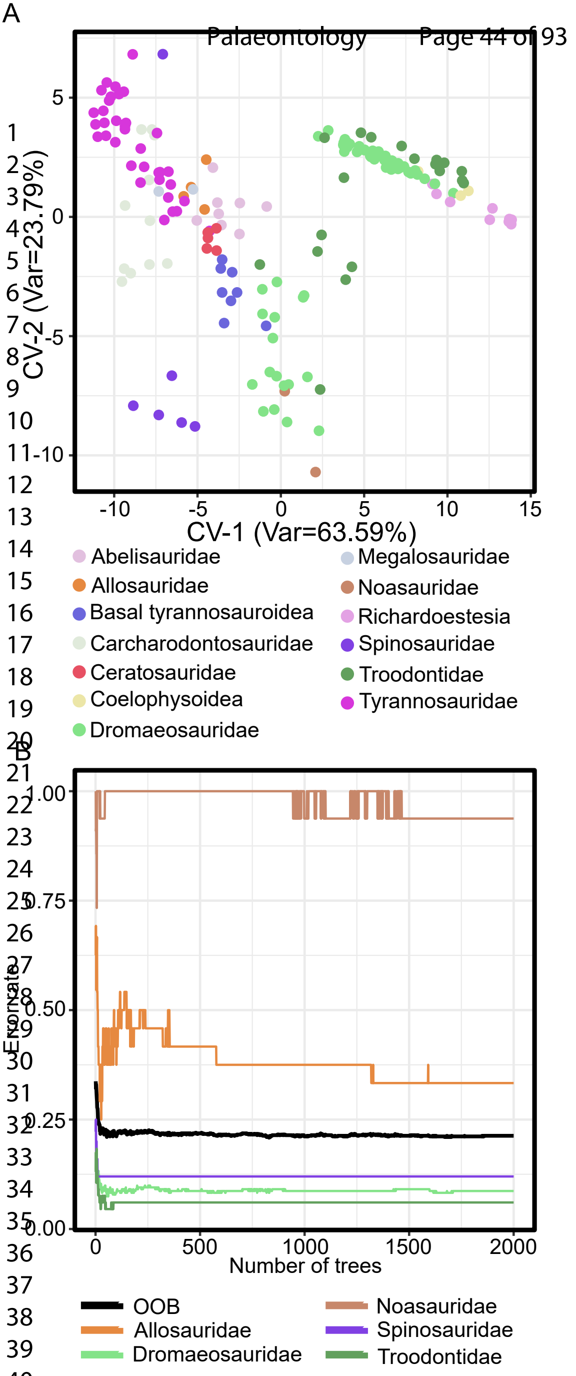
FIG. 14. Classification changes at the clade level using LR, MDA, RF classifiers and a majority vote ensemble classifier for the Larson *et al.* (2016) 17-class data. Vertical bars represent the clade predictions for each classifier, flows between the bars represent changes in prediction between the different classifiers. The ensemble classifier has an additional ‘unknown’ class where none of the individual classifiers were in agreement with a prediction. Pmx: pre-maxillary tooth.

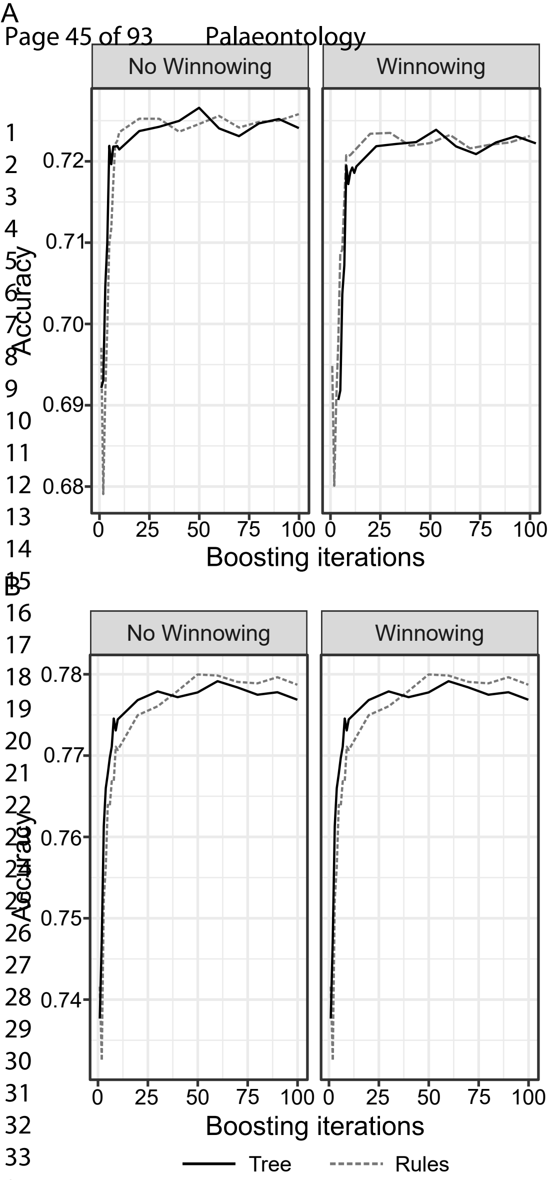

Final destination
Brighton
London
Wrong train
Station?
London Bridge
Victoria
Wrong train
Direct train?
No
Yes
Wrong train
Arrives between 08:00 and 08:15?
No
Yes
Wrong train
Correct train

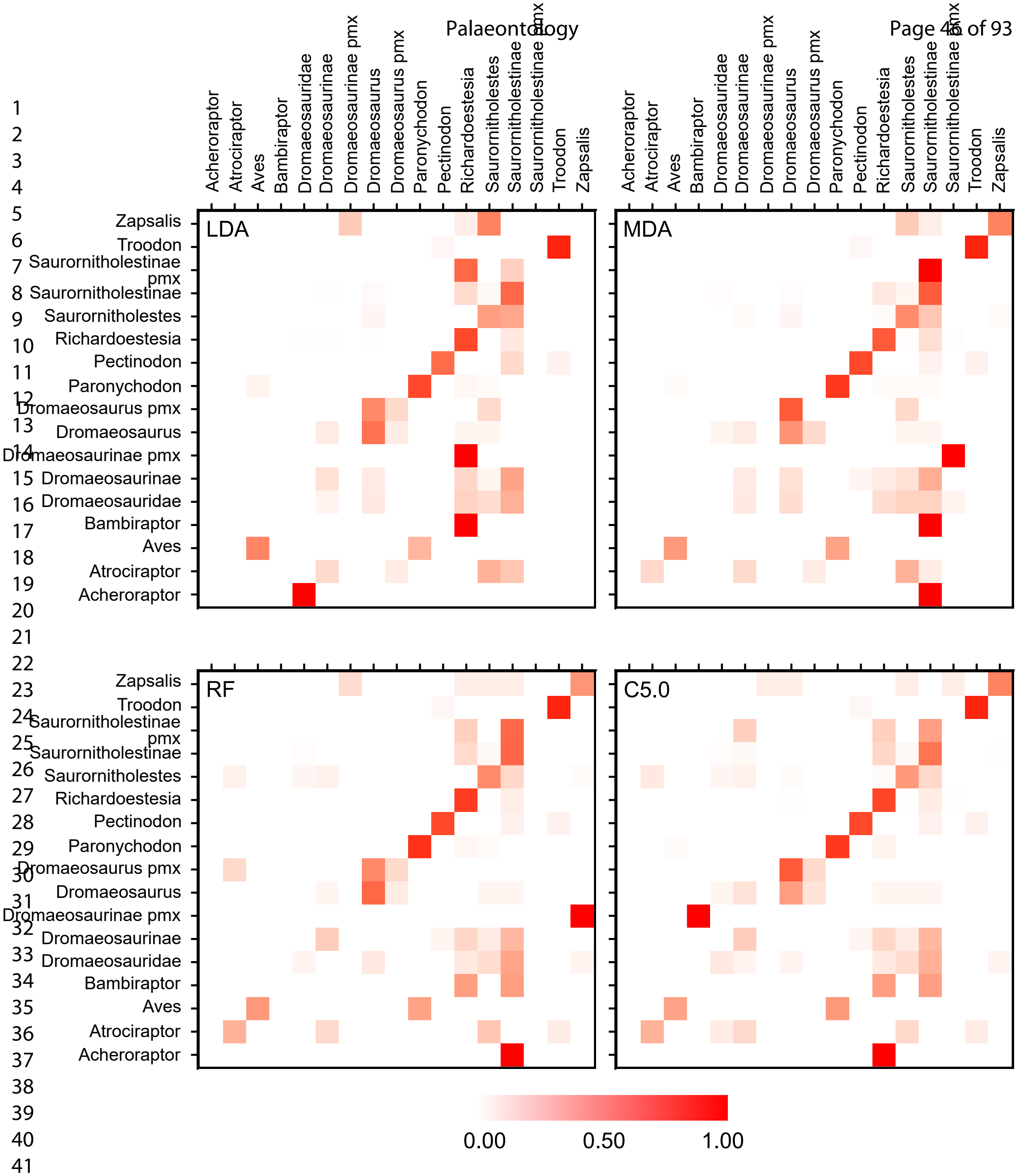

ADM
PDM
CH
CBL
CBW
CBL

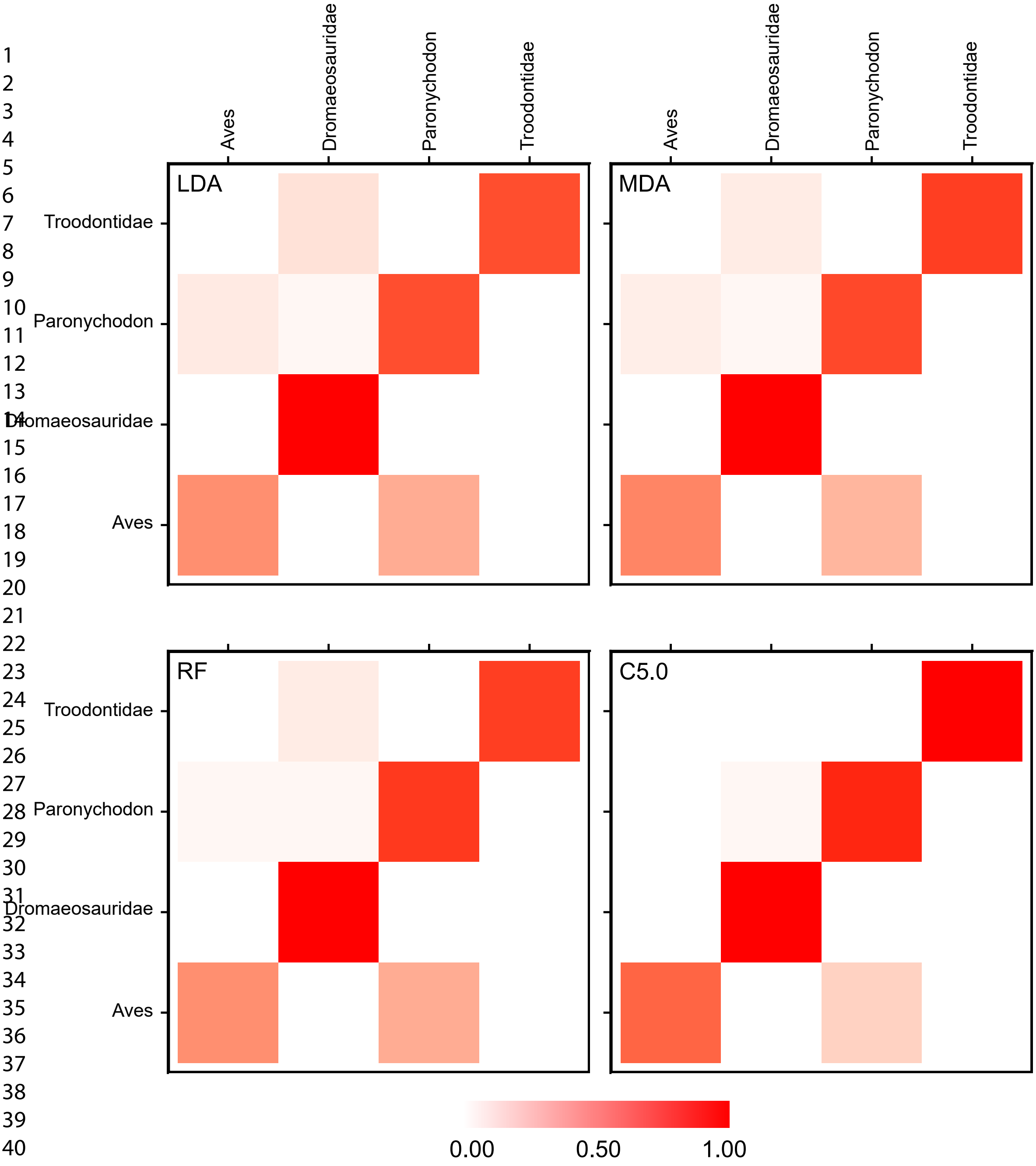

A
Training data.
New data sets with missing data inserted by random replacement of original training data.
C5.0 classifiers trained on each new data set.
Final predictions.
Each classifier produces an individual prediction.
B
Training data.
Data set with randomly inserted missing data.
New data sets with missing data replaced by imputed data. Here we imputed five times to create five new data sets from the original training data.
Primary classification.
C5.0 classifiers trained on each imputed data set.
Primary predictions.
Each classifier produces a prediction based on its imputed training data set.
Aggregation stage.
All the primary model predictions are combined.
A secondary C5.0 classifier is trained using the aggregated data as input.
Final predictions.

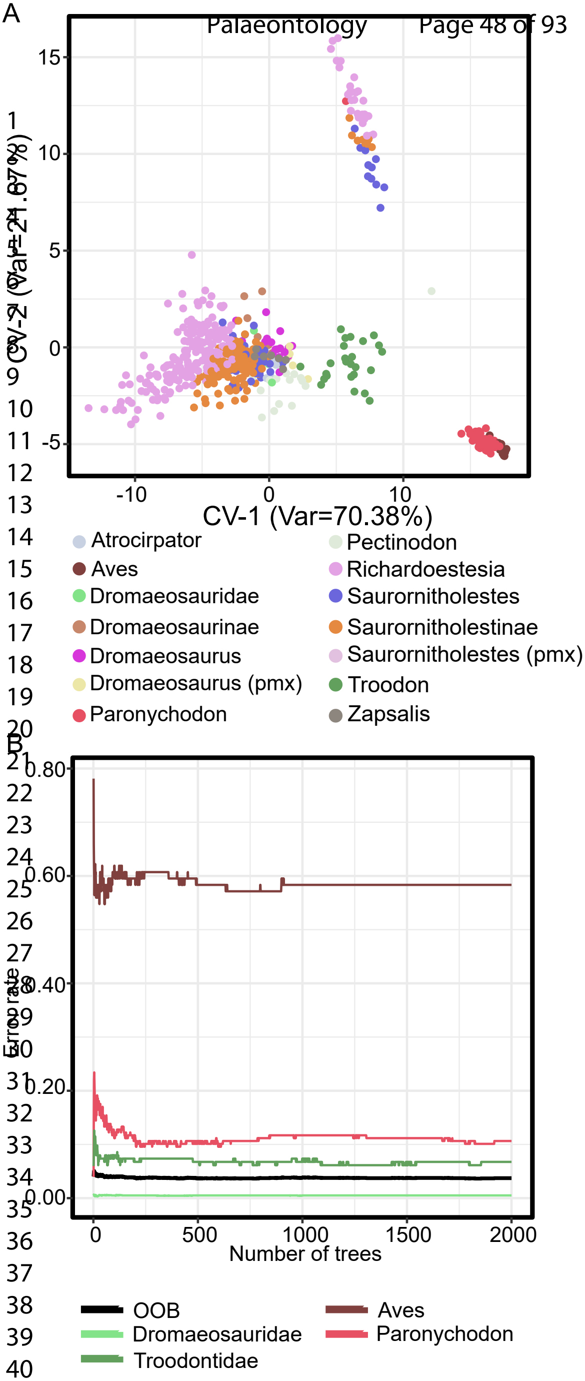

Tyrannosauridae
Troodontidae
Spinosauridae
Richardoestesia
Noasauridae
Megalosauridae
Dromaeosauridae
Coelophysoidea
Ceratosauridae
Carcharodontosauridae
Basal tyrannosauroidea
Basal theropoda
Allosauridae
Abelisauridae
LDA
MDA
RF
C5.0
0.00
0.50
1.00

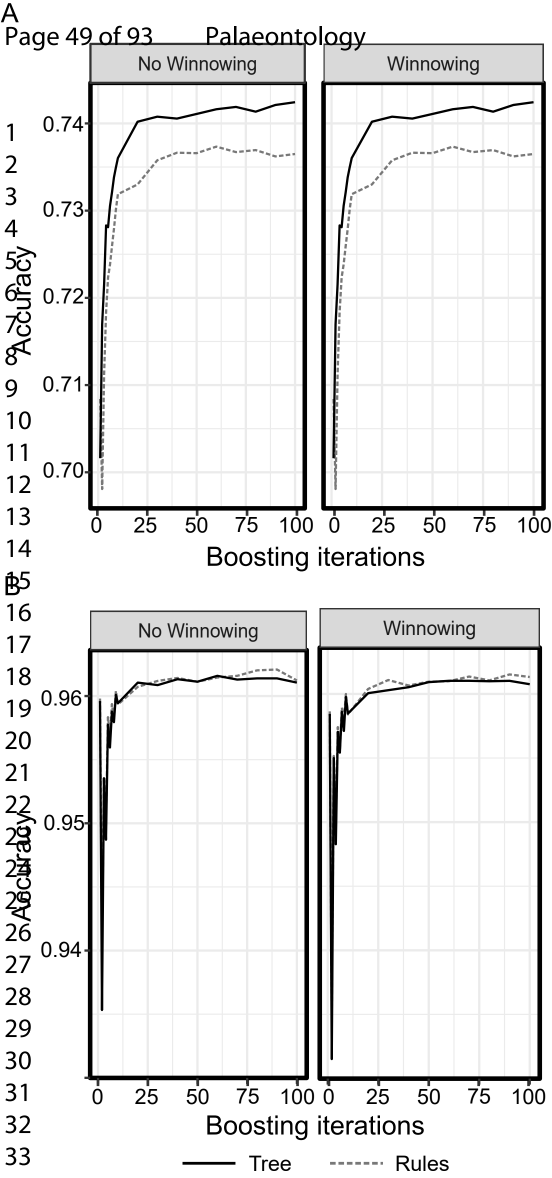

A
5
0
-5
-10
CV-2 (Var=23.79%)
-10
-5
0
5
10
15
CV-1 (Var=63.59%)
● Abelisauridae
● Allosauridae
● Basal tyrannosauroidea
● Carcharodontosauridae
● Ceratosauridae
● Coelophysoidea
● Dromaeosauridae
● Megalosauridae
● Noasauridae
● Richardoestesia
● Spinosauridae
● Troodontidae
● Tyrannosauridae
B
1.00
0.75
0.50
0.25
0.00
Error rate
0
500
1000
1500
2000
Number of trees
OOB
Allosauridae
Dromaeosauridae
Noasauridae
Spinosauridae
Troodontidae

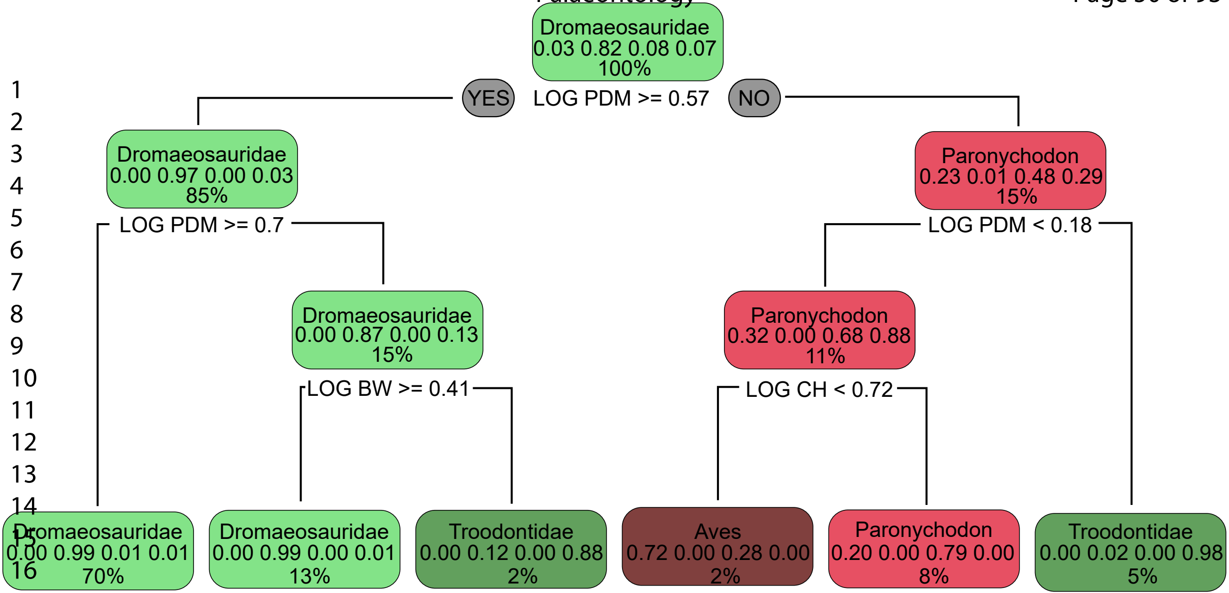

A
No Winnowing
Winnowing
Accuracy
0.72
0.71
0.70
0.69
0.68
0
25
50
75
100
Boosting iterations
B
No Winnowing
Winnowing
Accuracy
0.78
0.77
0.76
0.75
0.74
0
25
50
75
100
Boosting iterations
Tree
Rules


Acheroraptor
Atrociraptor
Aves
Bambiraptor
Dromaeosauridae
Dromaeosaurinae
Dromaeosaurinae pmx
Dromaeosaurus
Dromaeosaurus pmx
Paronychodon
Pectinodon
Richardoestesia
Saurornitholestes
Saurornitholestinae
Saurornitholestinae pmx
Troodon
Zapsalis
LDA
MDA
RF
C5.0
Zapsalis
Troodon
Saurornitholestinae pmx
Saurornitholestinae
Saurornitholestes
Richardoestesia
Pectinodon
Paronychodon
Dromaeosaurus pmx
Dromaeosaurus
Dromaeosaurinae pmx
Dromaeosaurinae
Dromaeosauridae
Bambiraptor
Aves
Atrociraptor
Acheroraptor
0.00
0.50
1.00


Aves
Dromaeosauridae
Paronychodon
Troodontidae
LDA
MDA
RF
C5.0
Troodontidae
Paronychodon
Dromaeosauridae
Aves
0.00
0.50
1.00


A
CV-2 (Var=21.67%)
15
10
5
0
-5
-10
0
10
CV-1 (Var=70.38%)
● Atrocirpator
● Aves
● Dromaeosauridae
● Dromaeosaurinae
● Dromaeosaurus
● Dromaeosaurus (pmx)
● Paronychodon
● Pectinodon
● Richardoestesia
● Saurornitholestes
● Saurornitholestinae
● Saurornitholestes (pmx)
● Troodon
● Zapsalis
B
Error rate
0.80
0.60
0.40
0.20
0.00
0
500
1000
1500
2000
Number of trees
OOB
Dromaeosauridae
Troodontidae
Aves
Paronychodon


A
No Winnowing
Winnowing
Accuracy
0.74
0.73
0.72
0.71
0.70
0
25
50
75
100
Boosting iterations
B
No Winnowing
Winnowing
Accuracy
0.96
0.95
0.94
0
25
50
75
100
Boosting iterations
Tree
Rules


Dromaeosauridae
0.03 0.82 0.08 0.07
100%
YES
LOG PDM >= 0.57
NO
Dromaeosauridae
0.00 0.97 0.00 0.03
85%
LOG PDM >= 0.7
Paronychodon
0.23 0.01 0.48 0.29
15%
LOG PDM < 0.18
Dromaeosauridae
0.00 0.87 0.00 0.13
15%
LOG BW >= 0.41
Paronychodon
0.32 0.00 0.68 0.88
11%
LOG CH < 0.72
Dromaeosauridae
0.00 0.99 0.01 0.01
70%
Dromaeosauridae
0.00 0.99 0.00 0.01
13%
Troodontidae
0.00 0.12 0.00 0.88
2%
Aves
0.72 0.00 0.28 0.00
2%
Paronychodon
0.20 0.00 0.79 0.00
8%
Troodontidae
0.00 0.02 0.00 0.98
5%

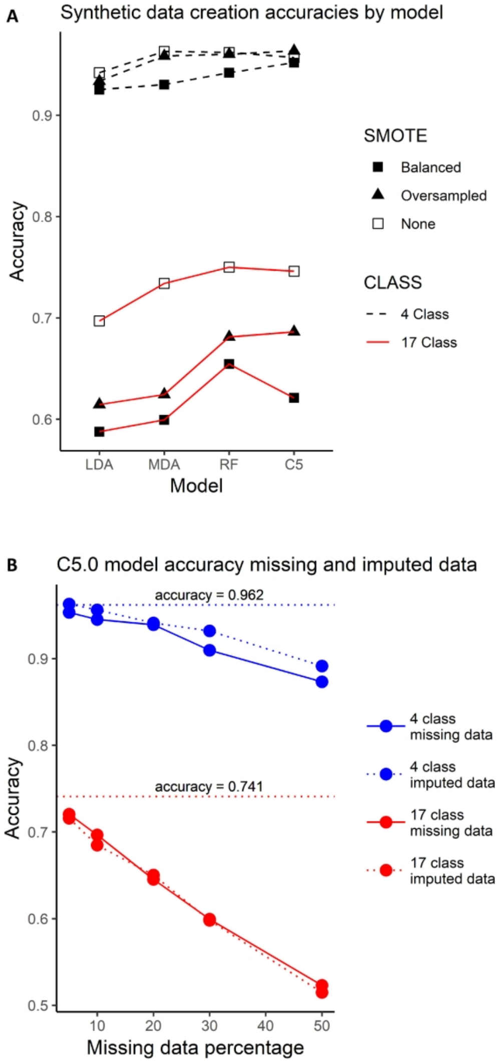
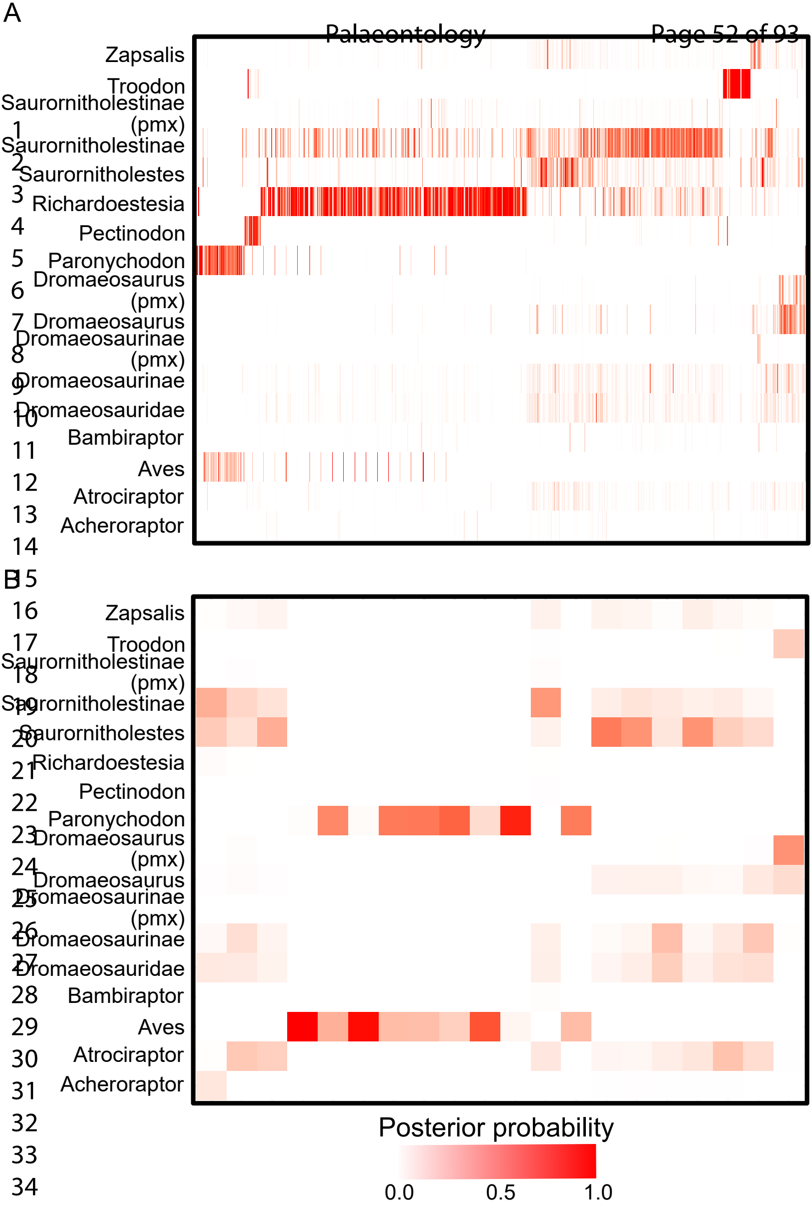
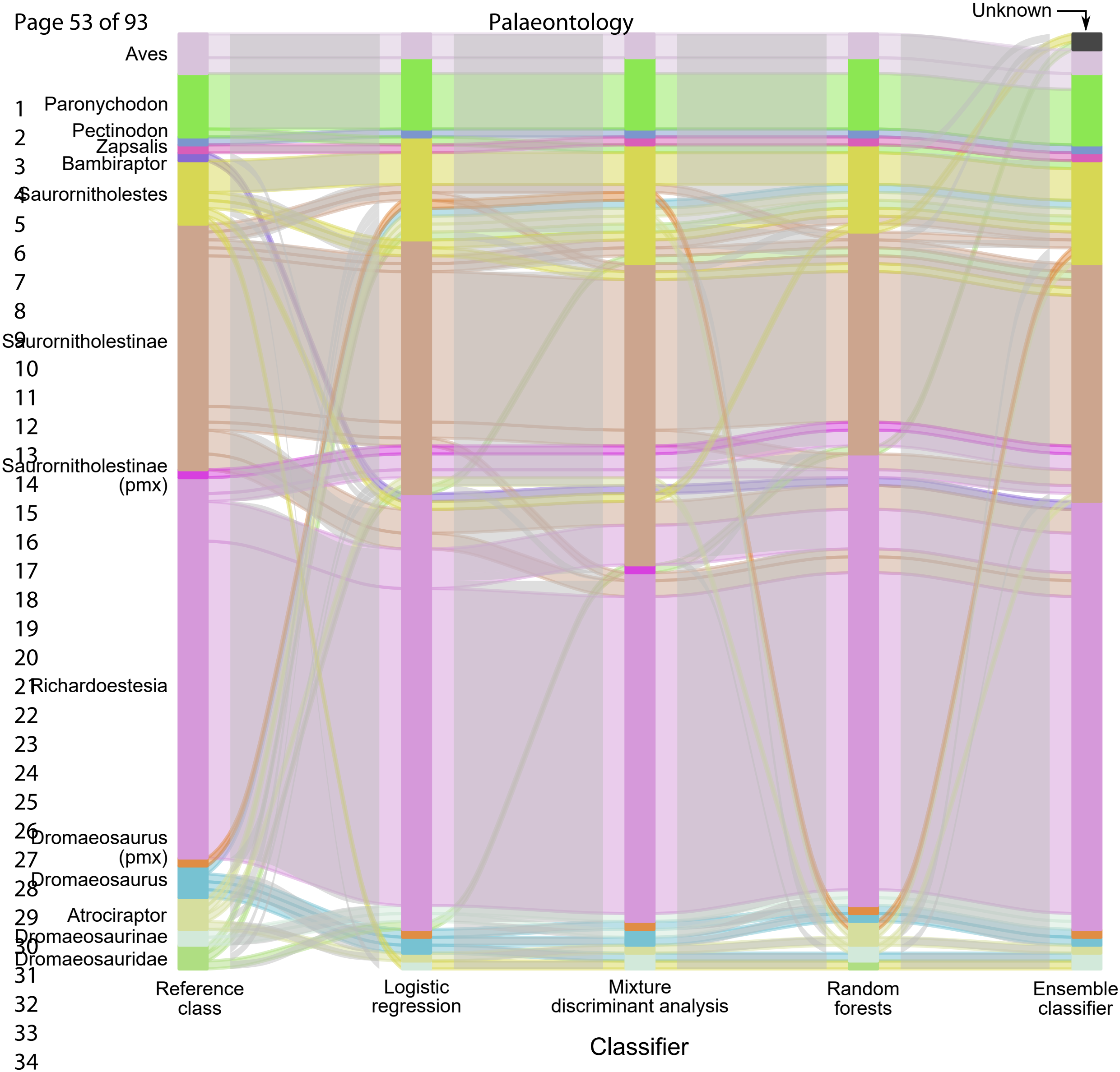


FIG. 12. A. C5.0 classifier accuracy for synthetically generated class balanced datasets B. C5.0 classifier accuracies for missing and imputed data at different levels. Horizontal dotted lines show the C5.0 model accuracy with no missing or imputed data.

209x296mm (300 x 300 DPI)


A
Zapsalis
Troodon
Saurornitholestinae (pmx)
Saurornitholestinae
Saurornitholestes
Richardoestesia
Pectinodon
Paronychodon
Dromaeosaurus (pmx)
Dromaeosaurus
Dromaeosaurinae (pmx)
Dromaeosaurinae
Dromaeosauridae
Bambiraptor
Aves
Atrociraptor
Acheroraptor
B
Zapsalis
Troodon
Saurornitholestinae (pmx)
Saurornitholestinae
Saurornitholestes
Richardoestesia
Pectinodon
Paronychodon
Dromaeosaurus (pmx)
Dromaeosaurus
Dromaeosaurinae (pmx)
Dromaeosaurinae
Dromaeosauridae
Bambiraptor
Aves
Atrociraptor
Acheroraptor
Posterior probability
0.0
0.5
1.0


Unknown
Aves
Paronychodon
Pectinodon
Zapsalis
Bambiraptor
Saurornitholestes
Saurornitholestinae
Saurornitholestinae (pmx)
Richardoestesia
Dromaeosaurus (pmx)
Dromaeosaurus
Atrociraptor
Dromaeosaurinae
Dromaeosauridae
Reference class
Logistic regression
Mixture discriminant analysis
Random forests
Ensemble classifier
Classifier

Term	Meaning	Reference(s)
Bagging	Also known as bootstrap aggregating. Used to reduce the variance of a decision tree classifier by creating training sample subsets on which to train the tree. A form of ensemble learning.	(Kuhn and Johnson 2013b)
Boosting	A process whereby many weak classifiers are combined into a strong classifier.	(Valiant 1984; Kuhn and Johnson 2013b)
C4.5	An algorithm used to create decision trees.	(Quinlan 1993; Salzberg 1994)
C5.0	An algorithm used to create decision trees. The successor to C4.5.	(Kuhn *et al.* 2018)
Decision trees	A supervised learning technique.	(Kuhn and Johnson 2013b)
Ensemble learning	Combining a group of classifier models to produce a final prediction.	(Hastie *et al.* 2009b)
Linear discriminant analysis (LDA)	A linear model for classification that seeks to find a combination of predictor values to categorise samples into groups. Also known as discriminant function analysis (DFA).	(Fisher 1936; Welch 1939)
Linear model	A model in which the terms that describe the model form a linear equation.	(Riffenburgh 2012)
Logistic regression (LR)	A linear model for regression and classification.	(Finch *et al.* 2014)
Machine learning	A method of data analysis in which the model learns from new data.	
Mixture discriminant analysis (MDA)	A non-linear extension to linear discriminant analysis.	(Hastie and Tibshirani 1996)
Naïve Bayes (NB)	A non-linear machine learning technique for group classification.	(Russell and Norvig 2009)
Non-linear model	A model in which the terms that describe the model do not form a linear equation.	(Riffenburgh 2012)
Posterior probability	The probability that a case can be assigned to a particular class after classification.	(Kuhn and Johnson 2013c)
Principle component analysis (PCA)	A technique to reduce the dimensionality of data whilst minimizing information loss.	(Jolliffe 2002)
Prior probability	In Bayesian statistics the prior distribution of the event i.e. the known or expected probability of an observation coming from a particular group before the classification is run.	(Kuhn and Johnson 2013c)
Pruning (winnowing)	A process to reduce overfitting of a model generated using the C5.0 algorithm.	(Kuhn *et al.* 2018)
Random forests (RF)	An algorithm used to create a series of uncorrelated decision trees which are combined into one model.	(Kuhn and Johnson 2013a)
Synthetic data	Data generated programmatically that does not exist in the original dataset.	

	Hendrickx, *et al.* (2015)				Larson, *et al.* (2016)			
	680 cases, 32 classes		886 cases, 14 classes		3020 cases, 17 classes		3020 cases, 4 classes	
	Accuracy							
	Model	Testing data	Model	Testing data	Model	Testing data	Model	Testing data
LDA	0.645	0.690	0.752	0.767	0.705	0.697	0.958	0.942
LR	0.687	0.730	0.753	0.759	0.721	0.726	0.962	0.951
MDA	0.745	0.774	0.803	0.796	0.732	0.734	0.965	0.963
NB	0.647	0.592	0.755	0.750	0.698	0.697	0.930	0.933
RF	0.742	0.758	0.786	0.804	0.748	0.750	0.965	0.962
C5.0	0.710	0.749	0.775	0.802	0.741	0.746	0.962	0.957

	17-class			4-class		
	Accuracy					
	Balanced	Oversampled	None	Balanced	Oversampled	None
LDA	0.588	0.614	0.697	0.925	0.934	0.942
MDA	0.599	0.624	0.734	0.930	0.958	0.963
RF	0.654	0.681	0.750	0.942	0.960	0.962
C5.0	0.621	0.686	0.746	0.952	0.963	0.957

	17-class		4-class	
	Accuracy			
	Missing data	Imputed data	Missing data	Imputed data
Percentage data missing / imputed				
0	0.741	0.741	0.962	0.962
5	0.721	0.716	0.953	0.963
10	0.696	0.685	0.945	0.956
20	0.645	0.650	0.939	0.941
30	0.599	0.598	0.909	0.932
50	0.523	0.515	0.873	0.891

	Hendrickx*, et al.* (2015) 14-class model		Larson*, et al.* (2016) 17-class model	
	Accuracy			
	equal priors	proportional priors	equal priors	proportional priors
Model				
LDA	0.767	0.774	0.697	0.708
MDA	0.796	0.841	0.734	0.746

	Case									
Taxon	1	2	3	4	5	6	7	8	9	10
Aves	0.57	0.00	0.00	0.00	0.98	0.02	0.00	0.00	0.00	0.00
Dromaeosaruidae	0.00	1.00	1.00	1.00	0.00	0.00	1.00	0.55	1.00	0.02
Paronychodon	0.42	0.00	0.00	0.00	0.02	0.98	0.00	0.00	0.00	0.00
Troodontidae	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.45	0.00	0.98

	Hendrickx*, et al.* (2015)		Larson*, et al.* (2016)	
	Accuracy			
	32 class	14 class	17 class	4 class
LR	0.685	0.733	0.731	0.958
MDA	0.749	0.785	0.733	0.963
RF	0.745	0.791	0.751	0.960
Ensemble stack	0.620	0.796	0.759	0.974
Majority vote	0.743	0.779	0.737	0.975

Comments from the Editor

- Reviewer 1 points out that the language and description of the study and in particular of the methodology is very technical and difficult to digest for non-experts. Where possible I suggest simplifying the description or providing a short, less technical summary to each step/section. In addition, you can think about adding a glossary to cover the most important technical terms.

> **Response:** We have added a glossary of terms to the Introduction (Table 1, new) and have simplified some of the technical descriptions – especially around the use of decision trees and random forests. The detail of this is included below in the specific responses to each reviewer. We have highlighted changes in red.

- Both reviewers would like to see more explanation and discussion with regards to the data set and its acquisitions. Reviewer 1 would like you to include R scripts for analysis in the dryad repository/supplementary information and include details of steps to perform the analysis.

> **Response:** We have added more information and discussion around the dataset (see below in the specific responses) and will upload R-scripts to dryad as requested.

- Please also see comments from the technical editor regarding formatting and figure quality. If your data set has been changed/adapted from the published source, think about whether it will be possible/appropriate to provide the data files used for the study.

> **Response:** We have uploaded the R scripts. As the data has not changed from the published source we have not provided this. All formatting and quality issues with the figures have been addressed (see below).

Comments from the technical editor

* Please respond directly to all referee comments, including the technical comments below. It is particularly important that you explain your reasoning if you have not followed any of the suggestions made in the reports.

> Response: Done

* Please upload your response to the reviewers as a separate document designated as a ‘Supplementary File’ with the other submission files. This will pull it into the automatically generated proof that is available to all reviewers.

> Response: Done

* If either of your co-authors has an ORCiD identifier, please ensure that this is included with their affiliation information on the manuscript. Ideally, we would prefer these to be linked through their ScholarOne accounts (as you have for the submitting author) so that their ORCiD accounts can be automatically updated with details of any published paper. However, we can add a link to the paper even without this.

> Response: Done

* If you re-order your reference list as part of your revision, or add any references, please take care to check any use of ditto marks (---). We do not actually need these at all in your submitted manuscript (full author names will be inserted and tagged, with ditto mark styling automatically added later as part of our production process). It does not matter whether you add or remove them from the existing list, but I would recommend including all names for any new reference.

Response: Done

* Referee 1 suggests restyling 'et al.' citations, but this would be done automatically as part of our production process.

Response: Noted

* Please supply all of your figures at a resolution of 600 dpi; preferably in tif format using LZW compression. Embedded photographs are fine at 300 dpi, but if any labelling is included the overall figure will require 600 dpi for printing. Please do not use jpg compression at any stage. Final widths should be either single column (80 mm), 2/3 page width (110 mm) or double column (166 mm). Please see the attached figure guidelines.

Response: Done. Please note that in order to clarify some of the text we have added one new figure (Fig. 1) and therefore have incremented the numbering on the figures listed below i.e. old Fig. 1 is now Fig. 2 etc. We have altered the caption numbering in the manuscript to reflect this. We have used pdf where appropriate to preserve resolution.

* Please view your figures at their final intended size on screen and check that all labels are in proportion and clearly legible. Generally, text sizes should be in a range equivalent to 6–10 pt Arial (viewed at 100%) although 6 pt size should not be used for critical text. However, this should be assessed by eye rather than relying on set font sizes as the absolute size of text will vary if the figure is resized. Please note that part labels (A, B, C…) on all figures are re-done by our typesetter to set them in a standard font at a height of 2 mm. It can look odd if other text is much larger than this.

* File size can also be reduced by using LZW compression and an 8-bit colour depth (VGA).

* Please confirm the final intended width of each figure (166 mm = full page; 110 mm = 2/3 page; or 80 mm = single column) by adding this to the file name (add 166, 110 or 80).

Response: Done

* Fig. 1: We will need a higher resolution original file for this image, even if it is intended to be set at single column width. At the current resolution, the text and arrows are clearly pixelated. Please do not scale up from this file as it will not result in a sufficient improvement in final quality. If you have an anti-aliasing option when exporting the image, this might help the appearance of the text and arrows.

Response: Done as pdf. This is now Fig. 2. A new Fig.1 is additionally supplied as pdf.

* Fig. 2: I would not recommend a landscape format for any image; it isn't convenient for readers either onscreen or on the page (especially if they are reading from a pdf format

article). The text on this image is too small for full page width, and I would recommend having it larger eve if this was intended for landscape format (although the supplied file is not high enough resolution for that). Can you redesign this into a portrait format by rotating all elements through 90 degrees? (And then increase the relative size of the text?)

Response: Redesigned as full page portrait and simplified. This is now Fig. 3

* Fig 3, 6: Again, this needs adjusting so that all text is clearly visible when the figure is viewed at 166 mm wide in a portrait format. Does setting all 4 images vertically help at all? The resolution of the file is slightly under 600 dpi at full page width.

Response: These are now Figs. 4 & 7. We have reset the individual elements of the figure and removed the in-figure text for clarity.

* Fig. 4: The layout of this figure works much better, but the axis labels (particularly the numbering, but also the labels and legend) on the lower part are too small. We will need a higher resolution file.

Response: This is now Fig 5. We have reset the text as requested.

* Fig. 5: We will need a higher resolution file to set at full page width.

Response: This is now Fig. 6. Supplied as pdf

* Fig. 7: The images work better at full page width, but the axis label text needs to be relatively larger. The resolution of the file is slightly under 600 dpi at full page width.

Response: This is now Fig.7. We have reset the text as requested. Supplied as pdf.

* Figs 8, 9: The text on these figures is much better proportioned at full page width, but we will need a higher resolution file.

Response: These are now Figs. 9 & 10. Reset and supplied as pdf

* Fig. 10: The resolution of this file is plenty for full page width, and the text is clear. Please use leading zeros for all decimal numbers (e.g. 0.03).

Response: This is now Fig. 11. Reset text as requested.

* Fig. 11:

Response: This is now Fig. 12. No changes

* Fig. 12:

Response: This is now Fig. 13. Reset for clarity.

* Fig. 13: this figure should work at full page width if you could increase the relative size of the text.

Response: This is now Fig. 14. Reset as requested.

Referee: 1 (Christophe Hendrickx)

In their paper "Learning to see the wood for the trees: machine learning, decision trees and the classification of isolated theropod teeth", Wills and colleagues introduce new techniques to identify isolated theropod teeth more accurately using quantitative data. According to these authors, machine learning and decision trees offer better alternatives over principle component (PCA) and discriminant analyses (DFA), which are the standard quantitative methodologies to identify theropod teeth. This is a very technical contribution to the world of dental identification and I admittedly got lost many times reading the text due to my limited knowledge of computer analysis (something probably shared with my colleagues working on theropod teeth).

Consequently, I am unable to comment on the technicality of the paper and the robustness of the methods they present. Given their expertise, I, however, have little doubt that the new approaches presented by Wills and colleagues are sound and should be used in the future in combination with those used by most authors (i.e., DFA, cluster and cladistic analyses). That being said, I think that the paper can be presented in a less technical way and should certainly provide additional information on how to use these new methods in a much clearer and straightforward way for novices like me. I would, therefore, recommend the publication of this work with moderate revision, urging the authors to consider the following points before resubmitting their MS.

1) As previously said, the main problem I have with this contribution is its technicality. Some sections, such as those explaining the random forests and C5.0 techniques, were particularly difficult to follow and after reading them several times, I am still not sure I understood them properly. I understand that these sections are needed but wonder if the authors can not make them less complex to read, or if they could not move any sections that are particularly technical to the appendices. The abstract, for instance, introduces many terms I have never heard before while this section should be written in a comprehensive way to anyone interested in the field (which is definitely my case). I would for instance suggest to precise what are the standard linear models (PCA and DFA for instance, something I understood later) and define in a brief way what are "machine learning", "decision tree" and "posterior probabilities". Otherwise, many readers will be lost from the very first lines of this paper.

Response: We have added a glossary of terms to the Introduction and have added the new section (copied below) as a more general introduction to decision trees and random forests as this seemed to be the area causing most difficulty:

"*Decision trees*

The final methodologies we explore are a departure from the standard linear or non-linear families of classification models. Both random forests and C5.0 are decision tree-based techniques that expand on the seminal work of Breiman *et al.* (1984), which introduced classification and regression trees.

Before exploring the detail of the two techniques it is useful to understand the basics of a decision tree. This is something we use in everyday life to make decisions based on a series of criteria. A simple example would be what train to catch to get to a certain destination preferably without changing stations. In order to get to this decision we effectively run through a series of steps, each step is a question and the answer to the question dictates a path that the decision can follow. A suitable decision tree for such a choice is shown in Figure 1. Every decision tree is a nested hierarchy of questions and answers (if … then statements). For the example of catching a train to London Victoria station, the following hypothetical decision tree (one of many possible trees) might be followed:

If the final destination of the train is Brighton, then it is the wrong train

or

If the final destination of the train is London, and the station is London Bridge, then it is the wrong train

or

If the final destination of the train is London, and the station is Victoria, and it is a not direct train, then it is the wrong train

or

If then final destination of the train is London, and the station is Victoria, and it is a direct train, and it arrives between 08:00 and 08:15, then it is the correct train.

A decision tree is essentially a flowchart of questions or rules that leads down a path to a prediction. Data is inputted into the root node of the tree. The decision tree algorithm then progressively divides the data into smaller and smaller groups based on the splitting criteria until the point at which the dataset can either be split no more or it reaches a rule that orders the splitting to finish. Decision trees can either be regression trees where the predicted outcome is a value (e.g., a house price) or classification trees where the predicted outcome is categorical (e.g., a taxon). The concepts of decision trees and random forests are similar. A decision tree is effectively built upon the entire dataset to produce one tree. A random forest combines many decision trees into a single model, where each of the trees in the model is generated on random subsets of observations and variables.

The major advantages of decision trees over techniques such as LDA or logistic regression are that: 1) they can accommodate missing data; 2) there is no need for the data to conform to a normal distribution, as the techniques are non-parametric; 3) outliers have little effect on the final classification as they will rarely define a splitting node; 4) they can use both categorical and numerical data as predictor variables; and 5) transformed predictor variables (e.g. log transforms) have no effect on the tree structure (Feldesman 2002). A drawback with decision tree methods is that of overfitting the data. This is when a tree is grown that perfectly predicts the classification pattern of the training data by defining terminal nodes (or leaves) that fit particular idiosyncrasies of the training process, i.e. that are relevant to that particular dataset only. Tree-based methods are also prone to bias if some classes dominate the data and care needs to be taken to account for this prior to fitting. ”

Random Forests. Random forests (RF) is an ensemble learning method where a large number of uncorrelated decision trees are aggregated to form a final classification (Breiman 2001). This final classification is based either on an average of all the individual tree estimates (for regression trees) or a simple majority vote (for classification trees). The decision trees are built by randomly selecting predictors and observations to create individual trees. This random selection process increases the diversity in the forest and leads to a more robust prediction. Random predictors (i.e. variables) are used at each split in the tree which de-correlate the trees forming the forest. The number of predictors used is controlled by a parameter setting (m_{try}) which Kuhn and Johnson (2013) and Breiman (2001) recommend setting to the square root of the number of predictors. Random forest classifications are sensitive to the number of trees used to build the forest with error rates reducing with increasing numbers of trees. Random forests tend to be stable and produce good predictive performance. However, they do have a number of disadvantages: even though some parameters are controllable, such as the number of trees or the number of predictors available at each split, the actual make up of each tree and therefore the forest is random and the forest itself (not the prediction) is less easy to interpret than a single decision tree; training a large number of trees can have higher computational overhead than a simple single decision tree."

2**)** **Because the main goal of this article is to introduce new techniques that the authors want people like me to use in the future, I strongly recommend them to provide, in the supplementary information, a modus operandi that precisely explains how to perform MDA, NB, RF and C5.0. This should include details on how the datamatrix has to be written and what steps to follow to get the final results. I am for instance surprised that the authors did not provide any files such as the R script and the different datamatrices used in their analysis. Please provide all these files and explain how they have to be used by authors who wish to use these new technique to identify theropod teeth. Likewise, the authors use two datasets to test the potential of each methodology to identify theropod teeth, but never explain what to do if one is interested in identifying a single theropod tooth.**

Response: We have uploaded sample R-scripts to dryad which can be used with any data. We have included simple instructions in these files to allow other researchers to run the models.

3) The paper has a whole section dedicated to missing data but do not seem to tackle other issues inherent to the theropod dentition, i.e., heterodonty and dental similarity in closely related species. What I wish to know is how MDA, NB, RF and C5.0 are going to perform with larger dataset on theropod crown measurements, which will include a larger number of taxa, and teeth from a wider distribution along the jaw (i.e., more mesial and lateral teeth, which are quite different morphometrically). Will it increase or decrease the success rate of their techniques? Likewise, is it more interesting for these new methods to include a larger number of measurement variables such as the extension of the mesial carina and the crown length and width at mid-crown? I always favored cladistic analysis based on a dentition-based datamatrix to identify isolated theropod teeth over any morphometric techniques using quantitative data mainly because I have the strong feeling that the more theropod crowns from a wider range of taxa and a wider distribution along the tooth row will be included in a dataset, the weaker any morphometric techniques will perform. It would be great if the authors could give their opinion on the matter in the discussion section.

Response: We have expanded the Discussion to address these points, with the addition of the following text:

> "Classification of isolated teeth in this manner will improve with better data, namely more cases per clade, to train the classifiers on. The careful addition of new measurement variables may also improve classification accuracies. As machine learning techniques have already been shown to be able to successfully classify taxa even with evolutionary convergence (e.g. Hoyal Cuthill et al., 2019) it is likely that even highly heterodont theropod clades and clades exhibiting dental morphological convergence could be accurately distinguished given the right amount of data and careful pre-processing of the data. It is probable that in some circumstances a combination of a dentition-based cladistic analysis and morphometric analysis may achieve the best results."

4) If the conclusions summarize relatively well the results of their evaluation of the different techniques to classify isolated teeth the best possible way, I m still confused on the best methods to apply on large sized datasets of theropod crown measurements, i.e., those that in the future will include a large range of taxa with a wider distribution along the tooth row. Can the authors state precisely in the conclusion what are the morphometric techniques they recommend, following what procedure, under what precise conditions, favoring what measurement variables, and using what taxon-level grouping (i.e., species, genus, "subfamily-level", "family-level", or "superfamily-level" taxa). This will summarize the core of their paper and provide the information that everyone in the field really wants to know.

Response: We have expanded the Discussion and Conclusions to address this.

> Addition to Discussion:
>
> "The taxon-level grouping that is chosen will have an impact on the overall accuracy of the model simply because this controls the number of cases per group which in turn impacts on the ability of the classifier to accurately describe that group. An attempt to classify at a species level where each species is described by (say) four individual teeth will be less accurate than a genus level classification where each genus is represented by several hundred teeth."
>
> Addition to Conclusion:
>
> "As a result of this study we would recommend the use of decision trees as an alternative approach to LDA. The final aim of the analysis should guide the choice of random forest or C5.0. If the goal is to predict the taxon that a tooth falls into then random forests are a good choice. If the aim is to classify and to be able to see how the classification is built within the tree structure then C5.0 should be used. In practice we would recommend corroboration of any results by checking predictions with another technique, preferably via the use of ensemble classifiers. The use of such techniques on isolated theropod teeth demonstrates that high levels of predictive taxonomic accuracy are possible from simple qualitative data as long as care is taken to understand the structure of the data in question and the assumptions that various techniques require."

I wish I could review in a better may the methodology followed and new techniques presented by the authors but my expertise simply prevent me to do so. I provided minor suggestions and corrections in a pdf of the MS. I mainly wish the authors to use, in some references, more recent works instead of papers published dozens of years ago. And I also believe that "xxx, et al. (year)" should be written "xxx et al. (year)" with no comma.

Response: We have revised the references where appropriate.

Line by line comments by reviewer 1 and our response to them can be found in the attached commented pdf file.

Referee: 2 (Jennifer Hoyal Cuthill)

I found this to be a well written, thorough and informative study testing some interesting methods and I support its publication with minor revisions to the text (detailed below).

My one qualm about the study, in response to which I would like to see a brief justification added to the introduction and/or discussion is in the basic approach to data selection. Why should anyone want to do machine learning on human-selected and presumably hand-measured (please clarify that point somewhere) morphometric data, when you could instead do machine learning and taxonomic classification directly on primary data such as photographs (e.g. Hoyal Cuthill et al 2019 Science advances 5.8 (2019): eaaw4967)?

I can think of a couple of possible reasons, for example you might want to use morphometric measurements where there was a very strong a priori reason for studying a particular variable/s, or where it was essential that you know exactly which variables had contributed (and to what extent) to the classification or if data sample sizes were very strongly limited (which doesn't seem to the case here as they seem to have quite a large sample of teeth), or as a first step in method-testing designed to make the process as comparable as possible to previous studies. However, these motivations seem to me in general like they would be pretty secondary given the enormous advantages of direct data analysis such as removing the necessity to have prior knowledge of informative variables and saving huge amount of human labour by fully automating the measurement and analysis process (Hoyal Cuthill et al 2019).

> Response: We have added a justification in the Methods and cited the relevant Hoyal Cuthill *et al*. (2019) paper.
>
> Addition to Material and Methods:
>
> "The datasets comprise human-selected and hand measured morphometric data rather than measurements derived from photographs or other digital sources of information (such as CT data) which are also used in machine learning classifications (e.g. Hoyal Cuthill *et al.*, 2019). As such, it is inevitable that some degree of error will be introduced into the measurement process. However, given that the classification of isolated theropod teeth is a common requirement in vertebrate palaeontology, and the currently available datasets are all hand measured morphometric data we feel there is value in applying such techniques to this data."

I note that I still think the authors have made a worthwhile contribution to method testing, but I think the MS would benefit from a brief, balanced explanation of when their methods might be useful and when they might not be.

> Response: We have expanded the discussion and conclusion sections to address this.
>
> Addition to Discussion:
>
> "The taxon-level grouping that is chosen will have an impact on the overall accuracy of the model simply because this controls the number of cases per group which in turn impacts on the ability of the classifier to accurately describe that group. An attempt to classify at a species level

where each species is described by (say) four individual teeth will be less accurate than a genus level classification where each genus is represented by several hundred teeth."

Addition to Conclusion:

"As a result of this study we would recommend the use of decision trees as an alternative approach to LDA. The final aim of the analysis should guide the choice of random forest or C5.0. If the goal is to predict the taxon that a tooth falls into then random forests are a good choice. If the aim is to classify and to be able to see how the classification is built within the tree structure then C5.0 should be used. In practice we would recommend corroboration of any results by checking predictions with another technique, preferably via the use of ensemble classifiers. The use of such techniques on isolated theropod teeth demonstrates that high levels of predictive taxonomic accuracy are possible from simple qualitative data as long as care is taken to understand the structure of the data in question and the assumptions that various techniques require."

1. There seem to be a very large number of figures, most of which could probably be moved to supplementary material should space be an issue.

Response: as there has been no objection from the technical editor we have left these in.

Line by line comments:

2. Abstract: Could you briefly state the sample sizes somewhere in the abstract e.g. x morphometric measurements, from y teeth, from z specimens of b species.

Response: Done as below in abstract.

We chose two published datasets comprising 886 teeth from 14 taxa, and 3020 teeth from 17 taxa each with five morphometric variables per tooth.

3. As I believe these morphometric data were taken from a published study could you note this in line 32 e.g. …published 'morphometric data'.

Response: Done

4. Please state/summarise the various method classification accuracies in the abstract.

Response: We have reworded part of the abstract (below) to add some information, but as the analyses were run over a wide range of scenarios and are summarised in Table 2 in the main text we do not think it appropriate to place all the results in the abstract.

"Our results suggest that machine learning and decision trees yield superior results over a wide range of data permutations with decision trees achieving accuracies of 96% in classifying test data in some cases."

5. Intro, p2, lines 8-21. In the citations of demonstrated uses of machine leaning for classification tasks, no mention of automated taxonomic classification by phenotype is made, although this is directly relevant to this study, and the authors may find it helpful to cite our neontological precedent here:
Hoyal Cuthill, Jennifer F. et al. "Deep learning on butterfly phenotypes tests evolution's oldest mathematical model." Science advances 5.8 (2019): eaaw4967).

Response: Done

“The use of non-linear analytical techniques that draw upon the rapidly expanding field of machine learning and decision trees has remained mostly unexplored with respect to characterizing fossil vertebrate morphology (Monson *et al.* 2018). By contrast, other disciplines have rapidly embraced machine learning techniques to undertake classification, prediction and various modelling tasks (Christin *et al.* 2019). Applications range from ecological modelling (Džeroski 2001; Cutler *et al.* 2007), population monitoring (Britzke *et al.* 2011), automated taxonomic classification by phenotype (Hoyal Cuthill et al., 2019), …”.

6. P5 line 3 ‘the algorithms employed in these analyses’ please list in brackets the specific algorithms referred to.

Response: rather than listing here (as we feel it breaks the flow), we have put a reference in to Table 1 which contains this information.

7. Lines 25-26 ‘pixel based data’ – photographs are a particularly obvious data choice for direct machine learning, is this what you mean by pixel-based data? It’s a slightly odd phrasing do you perhaps instead mean either photographs or secondary data generated from them? If so can you briefly unpack this.

Response: we have changed this to 'digital images' rather than pixel-based data or photographs.

“Although we have employed these techniques on fairly simple morphometric measurements, there is no reason why the techniques discussed below could not be employed on more complex morphological datasets such as 3D-shape data or digital images.”

8. P 7 line 19, ‘ingest’? Slightly unusual word usage, maybe [use] instead.

Response: Done

9. P 9 line 58 and following page line 27. Can you clarify whether you did species level classifications or any other level below genus for at least one of the dataset. In general, I would expect species to be a better level than genus for ML classification if the data are available (because it allows for the possibility of informative variation between species within a genus).

Response: added an explanation

We did not undertake a species-level analysis due to the lack of species-level data with enough complete cases. This has now been made clear in the text.

10. P 15 line 45 Please specify the taxonomic levels you refer to.

Response: Done

Using the Hendrick*, et al.* (2015) dataset we ran the classifiers at two taxonomic levels, the first a genus level with 32 classes and 680 cases and the second at a higher (family) taxonomic level with 14 classes and 886 cases.

11. P19 31 please briefly reiterate that you analyse morphometric measurements of teeth (i.e. not photographs or anything else) for any reader who goes straight to that section.

Response: Done

> "Our results demonstrate that the non-linear and machine learning techniques we applied to hand measured morphometric data of isolated theropod teeth classification consistently outperform LDA."

12. P 21 line 9 please clarify what you mean by character versus qualitative data here. I was under the impression from the methods that all the data used were continuous, quantitative measurements. Correct? So what do you mean by qualitative data?

> Response: This was a drafting error and qualitative should have read quantitative. We have corrected the relevant passage below to avoid confusion.
>
> "Recent studies such as Hendrickx *et al.* (2019) suggest that apomorphic character-based morphological data is potentially a more useful tool for distinguishing isolated theropod tooth crowns than the use of morphometric data."

13. Line 13-18 'We feel however that the careful application of machine learning techniques using the frameworks discussed in this study demonstrate that qualitative morphometric data can be a useful discriminator for the classification of isolated theropod teeth.' Please rephrase this subjective statement to an objective summary of your results e.g. something like: However, we show that [whatever sort of] morphometric measurements can discriminate isolated theropod teeth with taxonomic accuracy up to y%.

> Response: We have re-phrased this section as below
>
> "However, we show that the careful application of machine learning techniques using the frameworks discussed in this study demonstrate that continuous quantitative morphometric data can also discriminate isolated theropod teeth with taxonomic accuracy of up to 96% in the specific data we used."

14 .Line 25. I don't believe your study in itself justifies your statement that the methods used could likely cope with convergence so you should either cut this or you could make a more general statement that ML methods have been shown to be able to successfully classify taxa even with evolutionary convergence e.g. Hoyal Cuthill et al 2019 which demonstrates successful subspecies classification of butterflies with extensive mimicry – however we used a different ML method of deep learning on photographs.

> Response: We have re-phrased this section as below
>
> "Classification of isolated teeth in this manner will improve with better data, namely more cases per clade, to train the classifiers on. The careful addition of new measurement variables may also improve classification accuracies. As machine learning techniques have already been shown to be able to successfully classify taxa even with evolutionary convergence (e.g., Hoyal Cuthill et al., 2019) it is likely that even highly heterodont theropod clades and clades exhibiting dental morphological convergence could be accurately distinguished given the right amount of data and careful pre-processing of the data. It is probable that in some circumstances a combination of a dentition-based cladistic analysis and morphometric analysis may achieve the best results."

The following new references have been added to the manuscript and are highlighted in the revised version.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. 2009b. Ensemble learning. *In* HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (eds). *The elements of statistical learning*. Springer-Verlag, New York, 745 pp.

HENDRICKX, C., TSCHOPP, E. and EZCURRA, M. D. 2020. Taxonomic identification of isolated theropod teeth: The case of the shed tooth crown associated with Aerosteon (Theropoda: Megaraptora) and the dentition of Abelisauridae. Cretaceous Research, 108, 104312. doi: 10.1016/j.cretres.2019.104312

HOYAL CUTHILL, J. F., GUTTENBERG, N., LEDGER, S., CROWTHER, R. and HUERTAS, B. 2019. Deep learning on butterfly phenotypes tests evolution's oldest mathematical model. Science Advances, 5, eaaw4967. doi: 10.1126/sciadv.aaw4967

KUHN, M. and JOHNSON, K. 2013b. Classification trees and rule-based models. *In* KUHN, M. and JOHNSON, K. (eds). Applied predictive modelling. Springer-Verlag, New York, 600 pp.

KUHN, M. and JOHNSON, K. 2013c. Measuring performance in classificaiton models. In KUHN, M. and JOHNSON, K. (eds). Applied predictive modelling. Springer-Verlag, New York, 600 pp.

RIFFENBURGH, R. H. 2012. Chapter 19 - Modeling Concepts and Methods. In RIFFENBURGH, R. H. (ed.) Statistics in Medicine (Third Edition). Academic Press, San Diego, 690 pp.

SALZBERG, S. L. 1994. C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. Machine Learning, 16, 235–240.

VALIANT, L. 1984. A theory of the learnable. Communications of the ACM, 27, 1134–1142.