Sains Malaysiana 52(7)(2023): 1901-1914
http://doi.org/10.17576/jsm-2023-5207-01

RFE-Based Feature Selection to Improve Classification Accuracy for Morphometric Analysis of Craniodental Characters of House Rats

(Pemilihan Ciri Berasaskan RFE untuk Meningkatkan Ketepatan Pengelasan dalam Analisis Morfometri Sifat Kraniodental Tikus Rumah)

Aneesha Balachandran Pillay[1], Dharini Pathmanathan[1,*], Arpah Abu[2] & Hasmahzaiti Omar[2]

[1]*Institute of Mathematical Sciences, Faculty of Science, Universiti Malaya, 50603 Kuala Lumpur, Malaysia*
[2]*Institute of Biological Sciences, Faculty of Science, Universiti Malaya, 50603 Kuala Lumpur, Malaysia*

Received: 24 October 2022/Accepted: 26 June 2023

ABSTRACT

In conventional morphometrics, researchers often collect and analyze data using large numbers of morphometric features to study the shape variation among biological organisms. Feature selection is a fundamental tool in machine learning which is used to remove irrelevant and redundant features. Recursive feature elimination (RFE) is a popular feature selection technique that reduces data dimensionality and helps in selecting the subset of attributes based on predictor importance ranking. In this study, we perform RFE on the craniodental measurements of the *Rattus rattus* data to select the best feature subset for both males and females. We also performed a comparative study based on three machine learning algorithms such as Naïve Bayes, Random Forest, and Artificial Neural Network by using all features and the RFE-selected features to classify the *R. rattus* sample based on the age groups. Artificial Neural Network has shown to provide the best accuracy among these three predictive classification models.

Keywords: ANN, machine learning, naïve Bayes, recursive feature elimination, traditional morphometrics

ABSTRAK

Dalam morfometri konvensional, para penyelidik sering mengumpul dan menganalisis data dengan menggunakan bilangan ciri yang besar untuk mengkaji variasi bentuk antara organisma biologi. Pemilihan ciri memainkan peranan penting dalam pembelajaran mesin algorithma untuk mengeluarkan ciri-ciri yang tidak relevan dan berlebihan. Penghapusan ciri rekursif (RFE) merupakan kaedah pemilihan ciri terkenal yang boleh mengurangkan dimensi data dan juga boleh membantu memilih subset sifat berdasarkan kedudukan kepentingan peramal. Dalam kajian ini, kita menjalankan RFE pada ukuran kraniodental linear bagi data *Rattus rattus* untuk memilih subset ciri terbaik bagi kedua-dua tikus jantan dan betina. Kita telah menjalankan kajian perbandingan berdasarkan tiga algoritma pembelajaran mesin seperti Bayes Naif, Hutan Rawak dan Rangkaian Neural Tiruan menggunakan semua ciri dan ciri terpilih secara RFE untuk mengelaskan sampel *R. rattus* berdasarkan kumpulan umur. Setelah memantau hasil nilai ketepatan yang diperoleh bagi ketiga-tiga modal tersebut, Rangkaian Neural Tiruan telah terbukti memberi ketepatan yang terbaik antara ketiga-tiga model ini.

Kata kunci: ANN; Bayes naif; morfometri tradisi; pembelajaran mesin; penghapusan ciri rekursif

Introduction

Conventional morphometrics using linear distance measurements of skulls have proven to be powerful for identification, classification, and analysis of skull variability among biological organisms. However, some linear measurements used in these studies may contain irrelevant and redundant features which can affect the efficiency of learning models which may lead to performance degradation of unseen data (Li et al. 2017). Therefore, applying feature selection techniques to select a subset of relevant features to be applied into machine learning would improve the learning performance and construct better generalization models (Li et al. 2017).

The black rat, *Rattus rattus*, native to the Indian Peninsula is a known ubiquitous rodent pest. In Malaysia, studies related to *R. rattus* based on their craniodental measurements are still lacking especially on feature selection techniques to select the subset of the best morphometric features. This study aims to implement some machine learning approaches such as Naïve Bayes (NB), Random Forest (RF), and Artificial Neural Network (ANN) in aiding the feature selection process and achieving dimensionality reduction which can provide reliable age classification of *R. rattus* based on Muhammad Ikbal et al. (2019).

Many researchers conducted traditional morphometrics analysis using linear measurements which are directly obtained from biological organisms. These measurements are later analyzed using multivariate statistical approaches to identify the morphological variation among groups of individuals (Chuanromanee, Cohen & Ryan 2019). Principal component analysis (PCA) based on the pooled within-group variance-covariance matrix was applied as an approach to study metric dental variation of major human populations (Smith 1991). Separate analyses were conducted using raw measurements and C-scores to observe the possible effects of size and shape (Brace & Hunt 1990; Smith 1991) which showed distinct patterns by sex in certain regions. Applied the Traditional morphometric approach was also used to study the evolution of Exogyrinae oysters. PCA was performed to observe dimorphism between two species; cluster analysis is applied to evaluate morphological similarities between the oysters and cladistic analysis is used to assign time boundaries for the examined oyster members (Abdelhady & Elewa 2010)..

Eventually, researchers began to incorporate machine learning into morphometric studies for taxonomic classification. Commonly used classification methods include NB, RF and ANN. NB is a classification technique that can be applied for binary and multiclass classification problems. This classifier is based on the Bayes' theorem that assumes that the attributes in a dataset are conditionally independent, given the class. This classification algorithm estimates the posterior probability of each class, given an object by using information from sample data (Sammut & Webb 2010).

RF is a popular ensemble tree-based learning approach that has been used to solve classification and regression problems (Apao et al. 2020). This algorithm is known to select predictors based on the importance scores to exhibit good classification performance on multiclass datasets. RF easily handles categorical variables and identifies non-linear patterns in the data (Chaudhary, Kolhe & Kamal 2016).

ANN models are inspired from the human brain that works as a paradigm to perform computations in an effective and efficient manner (Mas & Flores 2008). Neural networks are based on the understanding of the biological nervous system. They are built on adaptable processing units to produce an output signal as functions of the sum of their weighted inputs and a certain threshold value (Wu 1992).

We aim to implement recursive feature elimination (RFE), a well-known feature selection technique to determine the best feature subset using the craniodental measurements dataset obtained from Muhammad Ikbal et al. (2019). RFE trains a model iteratively, ranking the features according to their importance scores and then removing the lowest ranking predictors (Darst, Malecki & Engelman 2018). NB, RF and ANN are applied as predictive classification models to compare models' performance to classify age groups of both male and female *R. rattus* using the RFE-selected variables and the results are tabulated.

MATERIALS AND METHODS

DESCRIPTION OF THE *RATTUS RATTUS* DATA

In this study, we use the male and female *R. rattus* cranial and mandible measurements i.e., 20 morphometric variables (Muhammad Ikbal et al. 2019). Figure 1 and Table 1 show the parts of measurements taken based on Musser and Newcomb (1983) and Musser, Lunde and Son (2009).

The linear measurements of the male and female rats were extracted from the original dataset based on their age classes. The three age classes are based on the molar wear stages. Stage C2: Cusps are still visible on all molars and the link between the first and second lobes of the upper M3 is very narrow; C3: The longitudinal link between the first and second lobes of the upper M3 is larger and generally wider; C4: Upper M3 displays nearly total fusion of the first and second lobes of the longitudinal link that is wide, but it remains visible on the other molar cusps. We then split the original dataset of the linear measurements and age classes for each sex using the 70/30 train/test split (70% of the whole dataset used for training, and 30% for testing) before fitting it to the RFE model to prevent overestimation of accuracy in the empirical analysis (Gholamy, Kreinovich & Kosheleva 2018).

RECURSIVE FEATURE ELIMINATION (RFE)

RFE is an effective feature selection method that initially uses the entire set of features to build the model. Since RF deals well with high dimensional data problems (Darst, Malecki & Engelman 2018), we applied this algorithm on each iteration of the RFE model using the *R. rattus* training data. RFE then effectively ranks the attributes according to their importance scores, eliminating the weak features iteratively until a desired number of top-ranked features are selected. Based on the accuracy of different attribute subset sizes obtained, we then choose the top performing features from the RFE model for each sex by referring to the RFE performance profile plots. These selected linear measurements of the training data and test data are scaled at unit variance before being fitted into three predictive classification models (Misra & Yadav 2020).

TABLE 1. Craniodental measurement of *R. rattus* based on

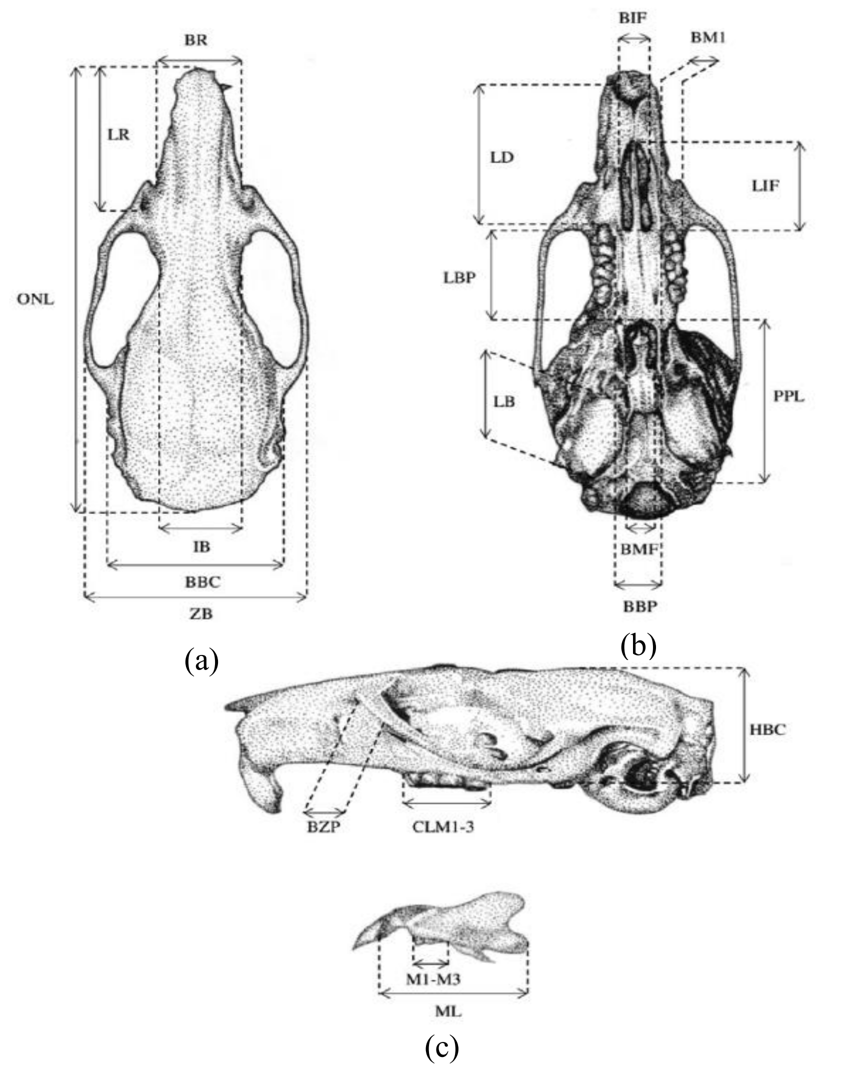


FIGURE 1. Craniodental measurements of *R. rattus* based on the (a) dorsal, (b) ventral, and (c) lateral views (Copyright: Muhammad Ikbal et al. (2019))

the dorsal, ventral, and lateral views

Dorsal	Ventral	Lateral	Mandible
Occipitonasal length (ONL)	Length of diastema (LD)	Breadth of zygomatic plate (BZP)	Length of mandible (ML)
Length of rostrum (LR)	Length of incisive foramina (LIF)	Crown length of maxillary molar row (CLM1.3)	Length of mandible toothrow (M1.M3)
Breadth of rostrum (BR)	Breadth of incisive foramina (BIF)	Height of braincase (HBC)	
Zygomatic breadth (ZB)	Breadth of first upper molar (BM1)		
Breadth of braincase (BBC)	Length of bony plate (LBP)		
Interorbital breadth (IB)	Length of auditory bulla (LB)		
	Post palatal length (PPL)		
	Breadth of mesopterygoid fossa (BMF)		
	Breadth across palate at first molar (BBP)		

CLASSIFICATION

Traditional morphometrics applies multivariate statistics to study morphological variations. In recent years, machine learning has been extensively used in morphometric studies for classification and identification tasks (Tan et al. 2018). We assess the performances of the NB, RF, and ANN methods on age classes of *R. rattus* males and females. We use the original feature set and the top performing features selected by the RFE model. A brief description of these predictive classification models is provided in Tan et al. (2018).

NAÏVE BAYES (NB)

The NB classification model is a classifier which provides a mechanism that utilizes predictors of the training data to estimate the posterior probability, $P(y_i \mid x_1, x_2, \ldots, x_n)$ (Sammut & Webb 2010). We trained the NB classifiers using all the scaled features as predictor variables and the age groups of *R. rattus* as class labels. This is done for both sexes and their performance measures are tabulated. The process is repeated for the RFE-selected features for comparison. Based on the *R. rattus* dataset, the Bayes theorem can be written as follows:

$$P(y_i \mid x_1, x_2, \ldots, x_n) = \frac{P(y_i)P(x_1, x_2, \ldots, x_n \mid y_i)}{P(x_1, x_2, \ldots, x_n)}$$

where $x_1, x_2, \ldots, x_n$ represents the scaled linear measurements and y_i represents the age classes (C2, C3 and C4) of the rats' training data.

RANDOM FOREST (RF)

RF has decision trees that train a dataset using the bootstrapping method. These decision trees reduce the chance of overfitting on the training data thus improving the predictive accuracy (Denisko & Hoffman 2018). We fitted the RF model with all the predictor variables of the training data using three age classes of rats as the classification category. The model is then assessed using the test data and the results of the performance measures are tabulated. The entire process is repeated using the training data with only the RFE-selected features for both sexes.

ARTIFICIAL NEURAL NETWORK (ANN)

ANN consists of several interconnected layers of information-processing units called neutrons and an input layer that processes the information of inputs. This information will be transferred to hidden layers. These layers process the information further before transferring it to the output layer which has one neuron that gives the function of the linear combination of the output obtained from the hidden layers (Bermejo et al. 2019). To fit the ANN model, we first input all features of the training data to the neural networks and select the age classes of rats as targets. This model is evaluated based on the results obtained by the confusion matrix. The process is repeated, by fitting only the RFE-selected features into the ANN model for both male and female *R. rattus* data.

EVALUATION METRICS OF MODELS' PERFORMANCE

We observed the multiclass confusion matrices of the classification models to compare their performances between the models with all features and models, with the selected features. The true positive (TP), true negative (TN), false negative (FN), false positive (FP) and accuracy (Acc) values after obtaining the confusion matrices are calculated. Since the target variable (age classes of rats) are imbalanced (Misra & Yadav 2020), i.e., 29.23% are C2, 23.08% are C3 and 47.69% are C4, we observed the Kappa, precision, recall and F1 score measures to evaluate the performance of the machine learning algorithms. These measures are calculated for each age class as follows:

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{Kappa} = \frac{\text{observed accuracy} - \text{expected accuracy}}{1 - \text{expected accuracy}}$$

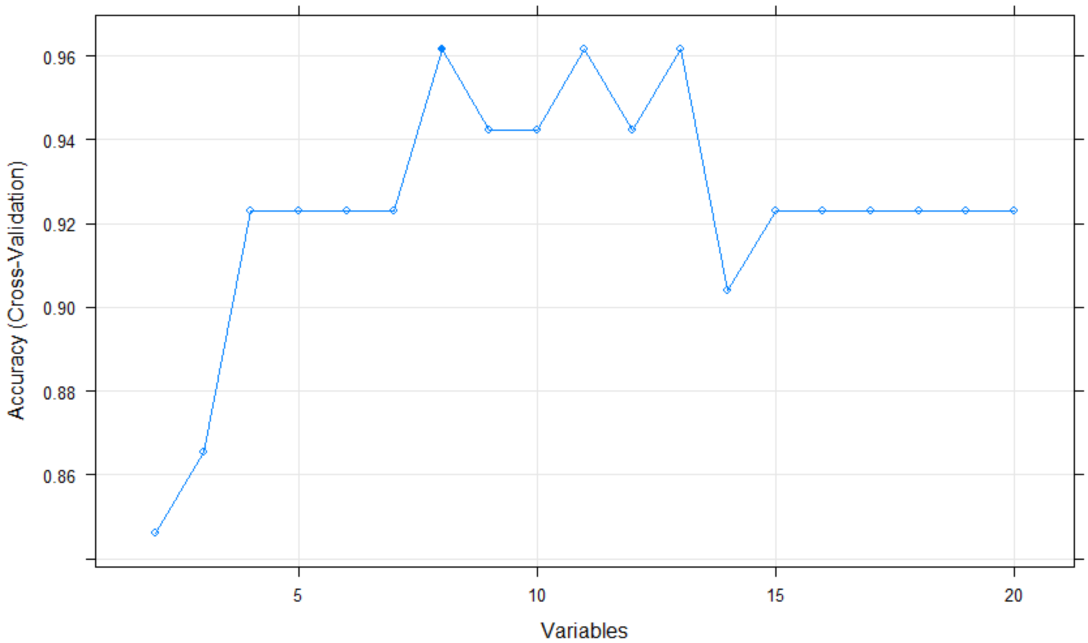
$$\text{F1 score} = 2 * \frac{Precision * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Receiver operating characteristic (ROC) curves are obtained. For the male and female rats, the respective Area under the ROC Curve (AUC) is obtained to assess the performance of the classification models with all features and models with RFE-selected features. The ROC curve plots the TP and FP, while the AUC calculates the area underneath the entire ROC curve which provides the overall measure of separability of age.

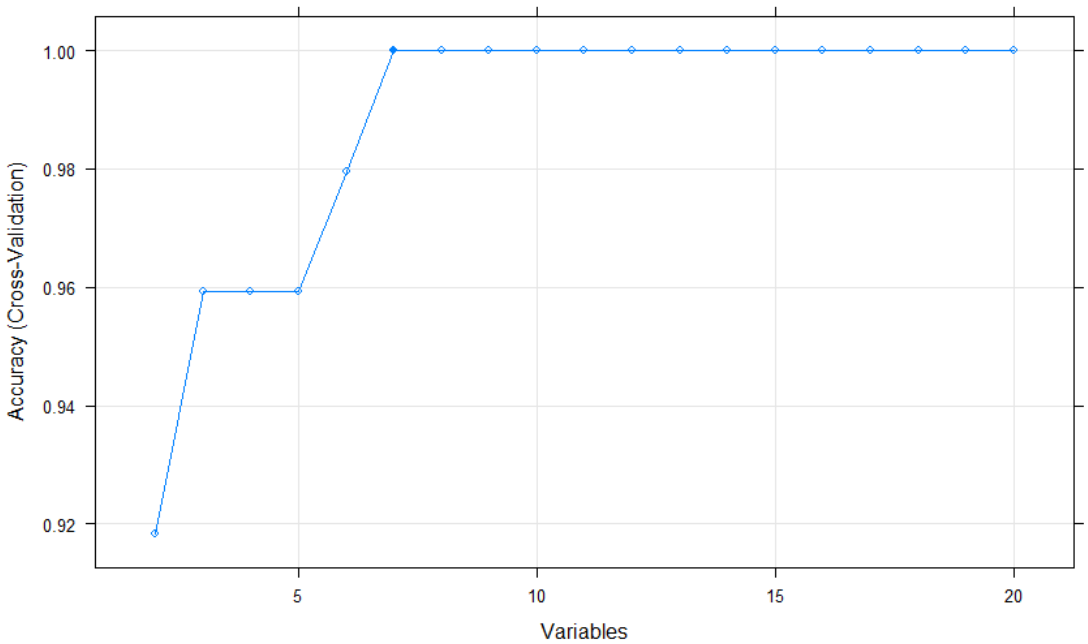
All statistical analyses were performed using R. We used the caret package (Kuhn 2008) in R to apply the RFE algorithm and to streamline the model training process for classification tasks. We also used the factoextra package (Kassambara & Mundt 2020) and ggfortify package (Tang, Horikoshi & Li 2016) in R to visualize the PCA output. The santaR package (Wolfer, Ebbels & Cheng 2022) is used to scale the linear measurements of both male and female rats at unit variance. The MLeval (John 2022) package is applied to construct the ROC curves for the classification models with all features and RFE-selected features.

RESULTS AND DISCUSSION

After performing the train-test split for the skull measurements data of both male and female rats, we applied the automatic RFE, by wrapping it around a random forest model to remove features recursively according to their age groups and the top performing features were selected. Based on the RFE results obtained in Figure 2, we selected HBC, IB, LD, BZP, BR, ZB, and LIF as the top performing features that indicate significant difference among age classes for the *R. rattus* males. These features may be selected by RFE as males of the *Rattus* genus are larger in size than females and can display a larger variation around the braincase compared to females. Male rats of this genus tend to have longer rostrum with shorter and wider zygomatic arch. As for *R. rattus* females, the top performing features to distinguish age classes are ZB, LD, BMF, BBC, IB and BR. These features are chosen using RFE as females of the *Rattus* genus display greater variation around the occipital bone with narrow zygomatic arch and longer magnum foramen (Alamoudi, Abdel-Rahman & Hassan 2021). All features selected by RFE for the male and female rats appear to coincide with most of the craniodental measurements used in Balakirev, Abramov and Rozhnov (2011), Breno, Leirs and Van Dongen (2011), Esselstyn et al. (2015), Libois et al. (1996), Motokawa, Lin and Lu (2004) and Timm et al. (2016). Among all features selected by RFE, we find that zygomatic breadth (ZB), interorbital breadth (IB), breadth of rostrum (BR), and length of diastema (LD) are significant in both male and female rats, which is 44.4% of the chosen features. These selected features are later used in the predictive classification models for each sex and their performance measures are evaluated and tabulated (Tables 2 & 3). Based on Tables 2 and 3, both training and test sets give excellent results in terms of accuracy for all three models, with all features included, where ANN is the best model for both male and female craniodental measurement datasets.



(a)



(b)

FIGURE 2. Performance profile plots across different subset sizes given by RFE approach for scaled (a) male and (b) female craniodental measurement dataset

TABLE 2. Model performance evaluation based on age groups for male *R. rattus*

Classifiers	Training data accuracy		Test data accuracy		Kappa	
	All features	Top performing features	All features	Top performing features	All features	Top performing features
NB	0.927	0.983	0.866	0.867	0.789	0.795
RF	0.943	0.963	0.933	0.800	0.891	0.685
ANN	0.980	0.987	1.000	1.000	1.000	1.000

TABLE 3. Model performance evaluation based on age groups for female *R. rattus*

Classifiers	Training data accuracy		Test data accuracy		Kappa	
	All features	Top performing features	All features	Top performing features	All features	Top performing features
NB	0.983	1.000	1.000	0.929	0.692	0.891
RF	1.000	0.975	1.000	0.857	0.841	0.781
ANN	1.000	0.994	1.000	1.000	1.000	1.000

We can observe a slight improvement in the NB model for *R. rattus* males after fitting the top performing features into the model. The lower test data accuracy for the RF model is due to the classification of the majority of the C4 age class as target variable in the male rats' test data. The overall performance evaluation of models with the top performing features for the male rats' data shows that ANN gives 100% test data accuracy and Kappa.

As for the female dataset, all three classification models for the top performing features show good results. Both training data and test data have accuracy of more than 97% for all models, with ANN being the best model. These results were further investigated using precision, recall and F1- score measures for the top performing features among both sexes (Table 4).

Based on the age groups of the male *R. rattus,* we observed that all the three models yield high scores for precision (Table 3). The recall measure shows good results for the models except for the C2 age group which for RF which is 0.5. This means that only half of the age class is correctly predicted. The F1 scores for all three models show that the groups are correctly identified and not disturbed by false results. The F1 score is considered perfect (1.000) for the ANN model for all male age groups. As for the age classes of the *R. rattus* females, all three models produce high scores for all the three measures. This indicates that the age classes are correctly classified based on the three models used.

All ROC-AUC curves (Figure 3) show promising results for all classification models with all features and top-selected features. There is an improvement in all the models when only the RFE-selected features were used. Based on the ROC-AUC curve for the female rats, all three classification models could clearly distinguish their age classes when only the top performing features were applied.

We choose the ANN as the best predictive classification model using the top five features for both the male and female rats based on the scores for all three measures considered and the ROC-AUC plots. Based on Figure 4, first two principal components (PCs) explained 95.2% of the variation in the age groups of male rats.

After considering the top performing features in the dataset, the first two PCs explain about 94.8% of the total variation. The clusters among age groups of male rats are more distinct when the RFE-selected features are used. As for the female *R. rattus*, the first two (PCs) explain 85.4% of the total variation in the age groups. When only the top performing features were considered, the PC1 and PC2 explain 93.8 % of the variation, which also shows more distinct clusters among the age groups in female rats. The increase in variation between age classes of both sexes based on the top performing features is because all PC components now contain only the eigenvectors of the features which are of importance. Based on the improvement shown in the PCA, selecting RFE-based features may also have more potential in examining the age variation of *R. rattus* using canonical variate analysis (CVA) and 'posteriori' Scheffe's test; a study conducted by Muhammad Ikbal et al. (2019) with 14 of the craniodental measurements for both sexes.

TABLE 4. Precision, recall and F1 scores of classification models using RFE-selected features for male and female *R. rattus* based on age groups

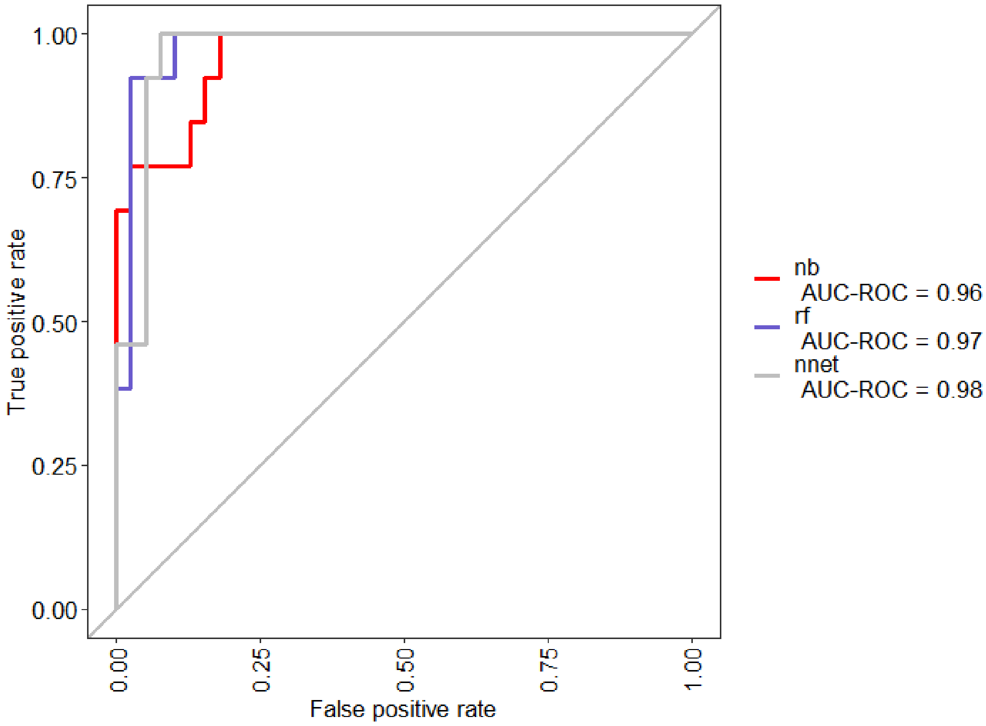
Male

Classification model	Precision			Recall			F1-score		
	C2	C3	C4	C2	C3	C4	C2	C3	C4
NB	1.000	1.000	0.750	1.000	0.600	1.000	1.000	0.750	0.857
RF	0.750	1.000	0.875	0.500	1.000	1.000	0.667	0.667	0.933
ANN	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

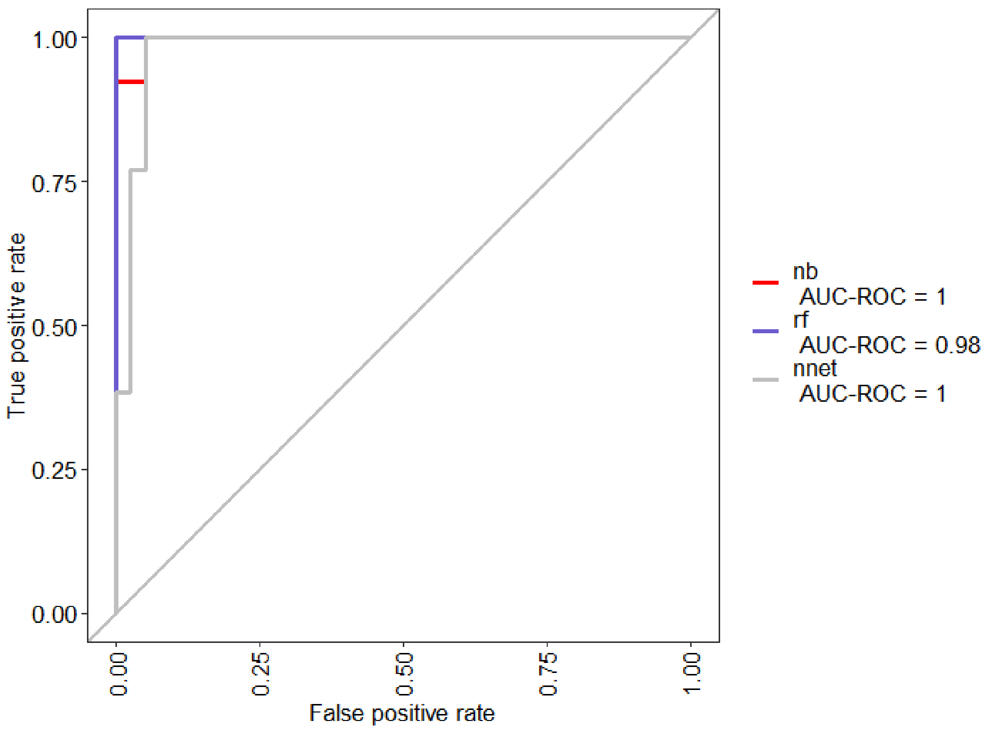
Female

Classification model	Precision			Recall			F1-score		
	C2	C3	C4	C2	C3	C4	C2	C3	C4
NB	1.000	0.750	1.000	0.833	1.000	1.000	0.909	0.857	1.000
RF	1.000	0.500	1.000	0.714	1.000	1.000	0.833	0.667	1.000
ANN	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

(a)

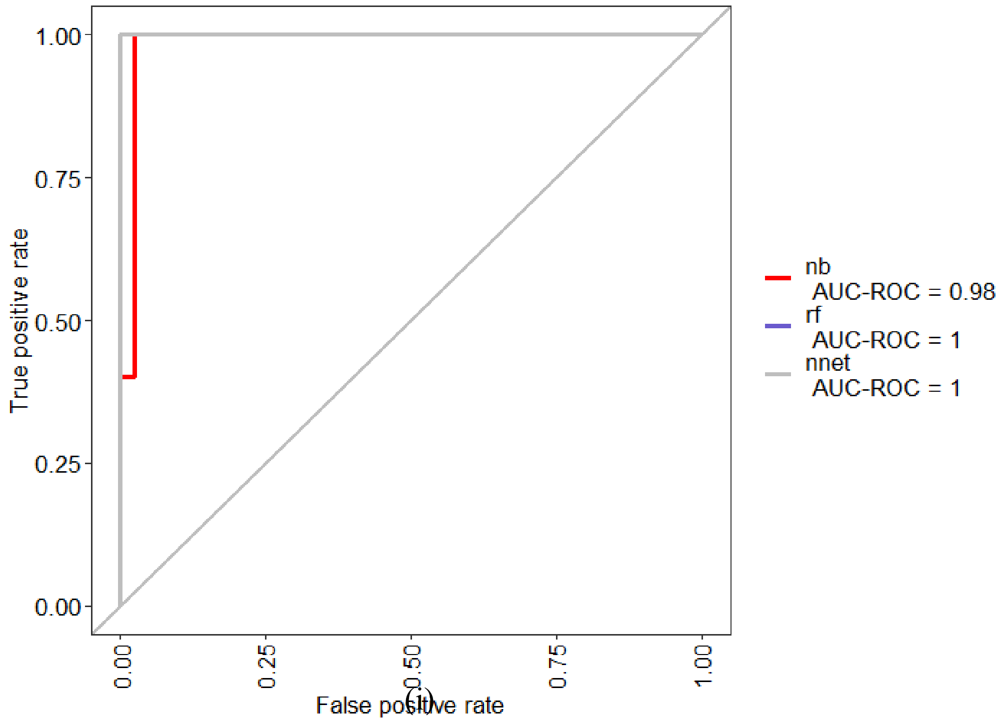

1.00
0.75
0.50
0.25
0.00
True positive rate
0.00
0.25
0.50
0.75
1.00
False positive rate
nb
AUC-ROC = 0.96
rf
AUC-ROC = 0.97
nnet
AUC-ROC = 0.98

(i)

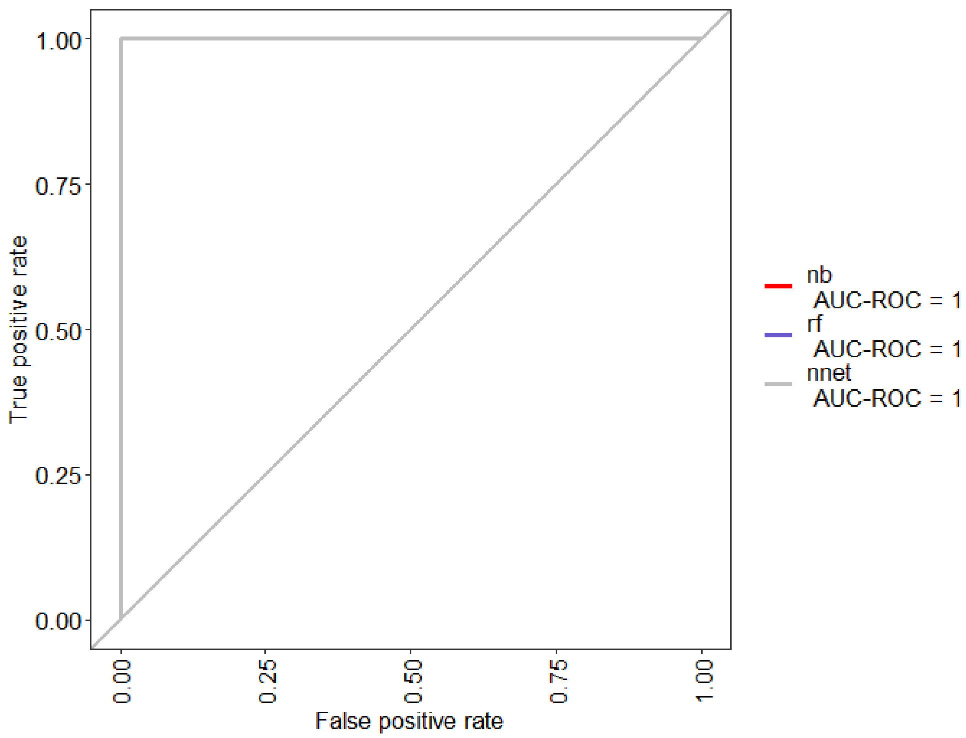

1.00
0.75
0.50
0.25
0.00
True positive rate
0.00
0.25
0.50
0.75
1.00
False positive rate
nb
AUC-ROC = 1
rf
AUC-ROC = 0.98
nnet
AUC-ROC = 1

(ii)

(b)



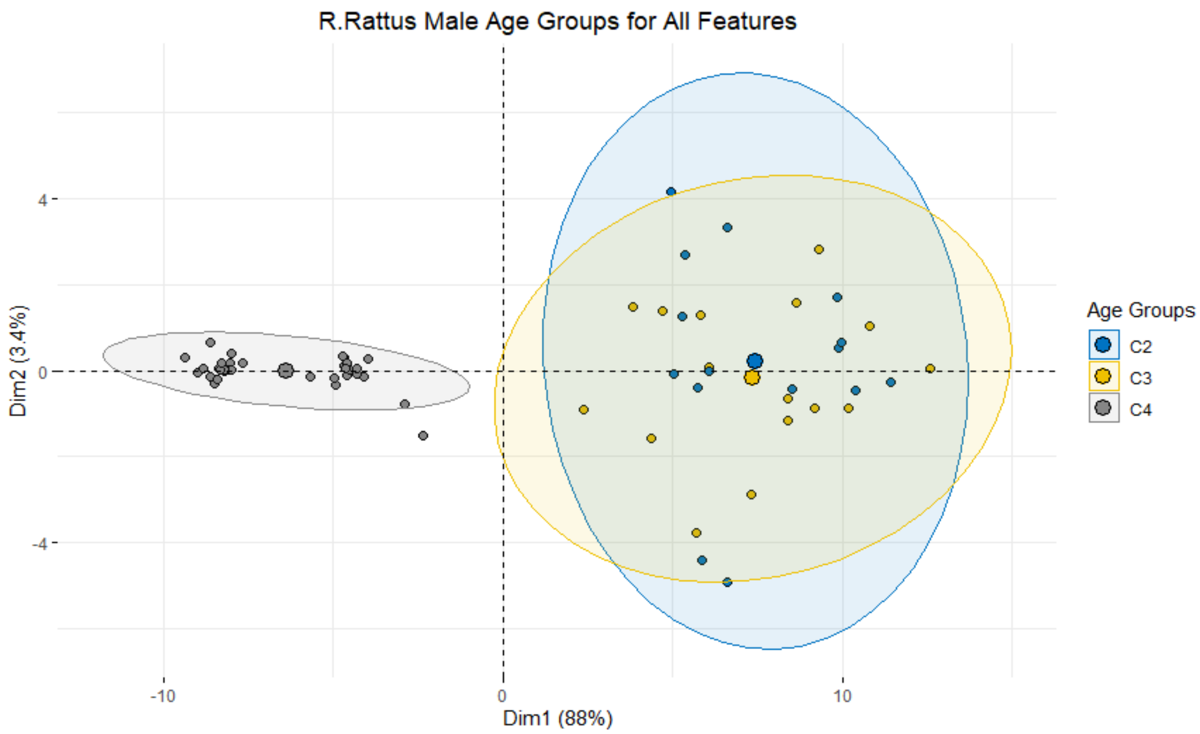
(i)



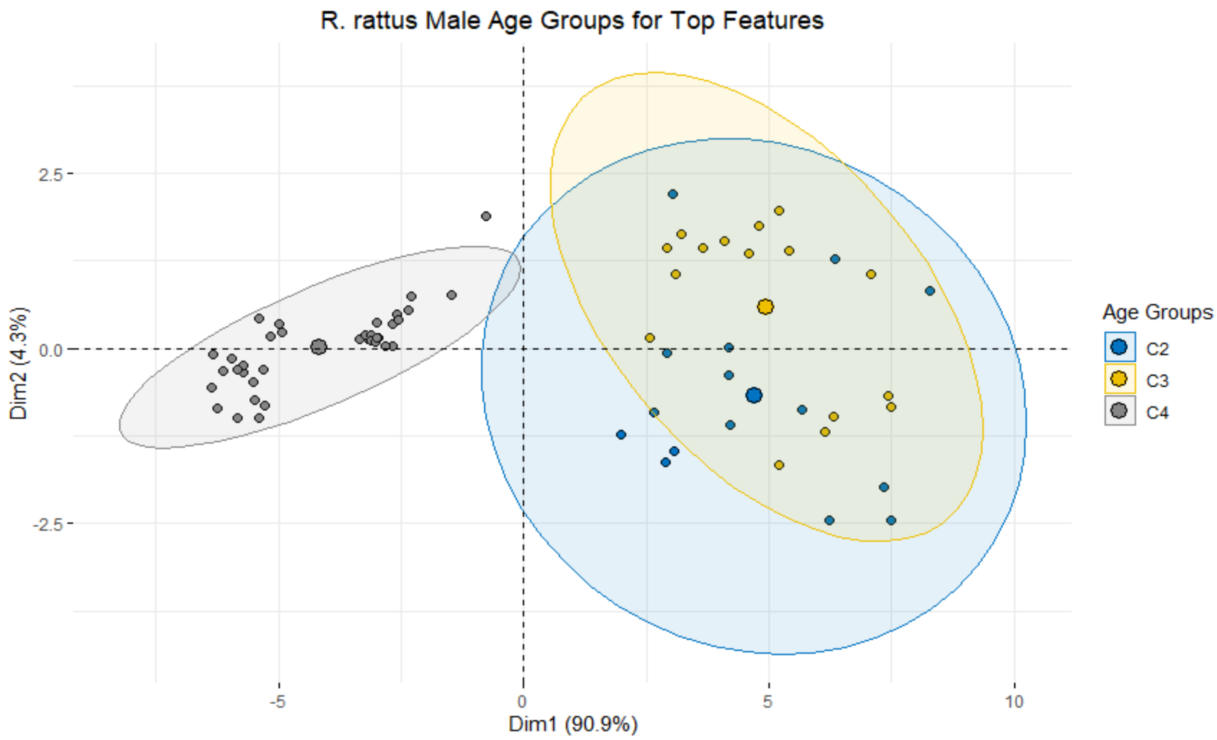
(ii)

FIGURE 3. ROC-AUC plots for *R. rattus* (a) male and (b) female craniodental measurement ((i) all features (ii) top performing features)

(a)

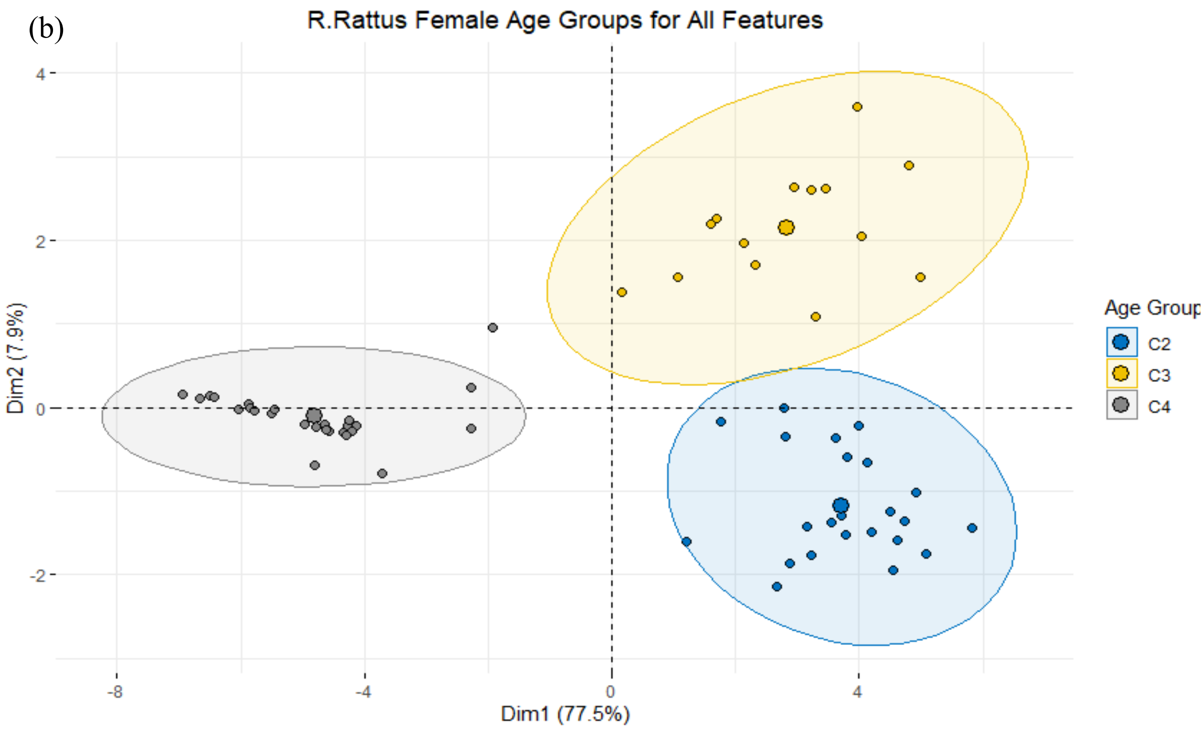

R.Rattus Male Age Groups for All Features
Dim2 (3.4%)
Dim1 (88%)
4
0
-4
-10
0
10
Age Groups
C2
C3
C4

(i)

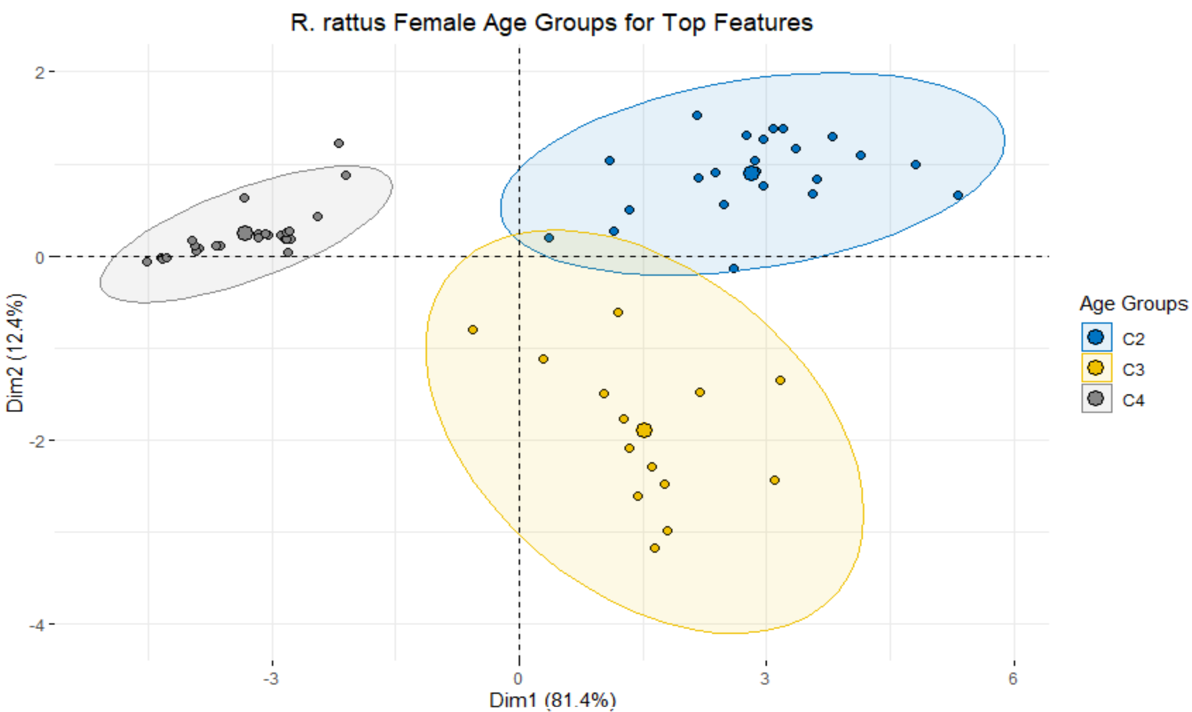

R. rattus Male Age Groups for Top Features
Dim2 (4.3%)
Dim1 (90.9%)
2.5
0.0
-2.5
-5
0
5
10
Age Groups
C2
C3
C4

(ii)

(b)



(i)



(ii)

FIGURE 4. PCA plots for *R. rattus* (a) male and (b) female craniodental measurement ((i) all features (ii) top performing features)

Conclusions

A good feature selection method that selects the best, highly discriminant features increase the performance of the model and reduces computational complexity in classification problems. Based on the analysis for *R. rattus* males and females, we noticed a slight improvement on the performance metrics of the three predictive classification models and in PCA when only the RFE-selected features are used. ANN outperforms the other models for both sexes. We also observed that using RFE as a feature selection method reduces computation complexity in morphometrics studies. Applying RFE-based features in the work done by Muhammad Ikbal et al. (2019) may achieve more promising results to observe the significance difference of *R. rattus* age groups and these features could also be used in other conventional morphometric studies of rats to examine their morphological differences. This study can also be extended by using the same approach for the classification of different biological organisms to produce a more generalized model and consider the effect of multicollinearity in traditional morphometric features when applying RFE to improve classification accuracy.

Acknowledgements

The authors thank the anonymous referees for their constructive comments. This work was supported by the Universiti Malaya, Faculty Research Grant (GPF088A-2020). The authors declare that they have no conflict of interests. The data that support the findings in this study are available from the corresponding author upon request.

References

Abdelhady, A.A. & Elewa, A.M.T. 2010. Evolution of the upper cretaceous oysters: Traditional morphometrics approach. In *Lecture Notes in Earth Sciences* 124: 157-176. Springer Verlag. https://doi.org/10.1007/978-3-540-95853-6_6

Alamoudi, M.O., Abdel-Rahman, E.H. & Hassan, S.S.M. 2021. Ontogenetic and sexual patterns in the cranial system of the brown rat (*Rattus norvegicus* Berkenhout, 1769) from Hai'l Region, Kingdom of Saudi Arabia. *Saudi Journal of Biological Sciences* 28(4): 2466-2475. https://doi.org/10.1016/j.sjbs.2021.01.048

Apao, N.J., Feliscuzo, L.S. Sta. Romana, C.L.C. & Tagaro, J. 2020. Multiclass classification using random forest algorithm to prognosticate the level of activity of patients with stroke. *International Journal of Scientific & Technology Research* 9: 1233-1240.

Balakirev, A.E., Abramov, A.V. & Rozhnov, V.V. 2011. Taxonomic revision of Niviventer (Rodentia, Muridae) from Vietnam: A morphological and molecular approach. *Russian Journal of Theriology* 10(1): 1-26. https://doi.org/10.15298/rusjtheriol.10.1.01

Bermejo, J.F., Juan F. Gómez Fernández, Fernando Olivencia Polo & Adolfo Crespo Márquez. 2019. A review of the use of artificial neural network models for energy and reliability prediction. A study of the solar PV, hydraulic and wind energy sources. *Applied Sciences (Switzerland)* 9(9): 1844. MDPI AG. https://doi.org/10.3390/app9091844

Brace, C.L. & Hunt, K.D. 1990. A nonracial craniofacial perspective on human variation: A(Ustralia) to Z(Uni). *American Journal of Physical Anthropology* 82(3): 341-360. https://doi.org/https://doi.org/10.1002/ajpa.1330820310

Breno, M., Leirs, H. & Van Dongen, S. 2011. Traditional and geometric morphometrics for studying skull morphology during growth in *Mastomys natalensis* (Rodentia: Muridae). *Journal of Mammalogy* 92(6): 1395-1406. https://doi.org/10.1644/10-MAMM-A-331.1

Chaudhary, A., Kolhe, S. & Kamal, R. 2016. An improved random forest classifier for multi-class classification. *Information Processing in Agriculture* 3(4): 215-222. https://doi.org/https://doi.org/10.1016/j.inpa.2016.08.002

Chuanromanee, T.S., Cohen, J.I. & Ryan, G.L. 2019. Morphological analysis of size and shape (MASS): An integrative software program for morphometric analyses of leaves. *Applications in Plant Sciences* 7(9): e11288. https://doi.org/10.1002/aps3.11288

Darst, B.F., Malecki, K.C. & Engelman, C.D. 2018. Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. *BMC Genetics* 19(1): 65. https://doi.org/10.1186/s12863-018-0633-8

Denisko, D. & Hoffman, M.M. 2018. Classification and interaction in random forests. *Proceedings of the National Academy of Sciences* 115(8): 1690-1692. https://doi.org/10.1073/pnas.1800256115

Esselstyn, J.A., Achmadi, A.S., Handika, H. & Rowe, K.C. 2015. A hog-nosed shrew rat (Rodentia: Muridae) from Sulawesi Island, Indonesia. *Journal of Mammalogy* 96(5): 895-907. https://doi.org/10.1093/jmammal/gyv093

Gholamy, A., Kreinovich, V., & Kosheleva, O. 2018. Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation. Departmental Technical Reports (CS). 1209. https://scholarworks.utep.edu/cs_techrep/1209

John, C.R. 2022. Package 'MLeval'. Machine Learning Model Evaluation.

Kassambara, A. & Mundt, F. 2020. Extract and visualize the results of multivariate data analyses [R Package Factoextra Version 1.0.7].

Kuhn, M. 2008. Building predictive models in *R* using the Caret package. *Journal of Statistical Software* 28(5). https://doi.

org/10.18637/jss.v028.i05
Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R.P., Tang, J. & Liu, H. 2017. Feature selection: A data perspective. *ACM Computing Surveys* 50(6): 94. https://doi.org/10.1145/3136625
Libois, R., Ramalhinho, G., da Luz Mathias, M., Santos-Reis, M., Fons, R., Petrucci-Fonseca, F., Oom, M. & Collares-Peirera, M. 1996. First approach on the skull morphology of the black rat (*Rattus rattus*) from Terceira and São-Miguel Islands (Azores Archipelago). *Vie et Milieu* 46(September): 245-251.
Mas, J.F. & Flores, J.J. 2008. The application of artificial neural networks to the analysis of remotely sensed data. *International Journal of Remote Sensing*. Taylor and Francis Ltd. https://doi.org/10.1080/01431160701352154
Misra, P. & Yadav, A.S. 2020. Improving the classification accuracy using recursive feature elimination with cross-validation. *International Journal on Emerging Technologies* 11(3): 659-665.
Mohamad Ikbal, Nurul Huda, Dharini Pathmanathan, Subha Bhassu, Khanom Simarani & Hasmahzaiti Omar. 2019. Morphometric analysis of craniodental characters of the house rat, *Rattus rattus* (Rodentia: Muridae) in Peninsular Malaysia. *Sains Malaysiana* 48(10): 2103-2111. https://doi.org/10.17576/jsm-2019-4810-05
Motokawa, M., Lin, L-K. & Lu, K-H. 2004. Geographic variation in cranial features of the polynesian rat *Rattus exulans* (Peale, 1848) (Mammalia: Rodentia: muridae). *The Raffles Bulletin of Zoology* 52(2): 653-663.
Musser, G.G. & Newcomb, C. 1983. Malaysian Murids and the Giant Rat of Sumatra. *Bulletin of the American Museum Natural History* 174: Article 4.
Musser, G., Lunde, D. & Son, N. 2009. Description of a new genus and species of rodent (Murinae, Muridae, Rodentia) from the Tower Karst Region of Northeastern Vietnam. *American Museum Novitates* 3517(September): 1-41. https://doi.org/10.1206/0003-0082(2006)3517[1:DOANGA]2.0.CO;2
R Core Team. 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
Sammut, C. & Webb, G. 2010. *Encyclopedia of Machine Learning*. Boston: Springer US. https://doi.org/10.1007/978-0-387-30164-8
Smith, F.H. 1991. Skull shapes and the map: Craniometric analysis in the dispersion of modern homo. By W.W. Howells, Vol. 79, Papers of the Peabody Museum of Archaeology and Ethnology. Cambridge: Harvard University Press. 1989. 187. *American Journal of Physical Anthropology* 86(1): 89-90. https://doi.org/https://doi.org/10.1002/ajpa.1330860110
Tan, J., Chang, S-W., Abdul Kareem, S., Yap, H.J. & Thai, Y-K. 2018. Deep learning for plant species classification using leaf vein morphometric. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 17(1): 82-90. https://doi.org/10.1109/TCBB.2018.2848653.
Tang, Y., Horikoshi, M. & Li, W. 2016. Ggfortify: Unified interface to visualize statistical results of popular R packages. *The R Journal* 8(2): 474. https://doi.org/10.32614/RJ-2016-060
Timm, R.M., Weijola, V., Aplin, K.P., Donnellan, S.C., Flannery, T.F., Thomson, V. & Pine, R.H. 2016. A new species of Rattus (Rodentia: Muridae) from Manus Island, Papua New Guinea. *Journal of Mammalogy* 97(3): 861-878. https://doi.org/10.1093/jmammal/gyw034
Wolfer, A., Ebbels, T. & Cheng, J. 2022. Package 'SantaR'. Short Asynchronous Time-Series Analysis.
Wu, B. 1992. An introduction to neural networks and their applications in manufacturing. *Journal of Intelligent Manufacturing* 3(6): 391-403. https://doi.org/10.1007/BF01473534

*Corresponding author; email: dharini@um.edu.my