ON SOME METHODS OF FEATURE ENGINEERING USEFUL FOR CRANIODENTAL MORPHOMETRICS OF RATS, SHREWS AND KANGAROOS

ANEESHA PILLAY A/P BALACHANDRAN PILLAY

FACULTY OF SCIENCE
UNIVERSITI MALAYA
KUALA LUMPUR

2024

ON SOME METHODS OF FEATURE ENGINEERING USEFUL FOR CRANIODENTAL MORPHOMETRICS OF RATS, SHREWS AND KANGAROOS

ANEESHA PILLAY A/P BALACHANDRAN PILLAY

THESIS SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

INSTITUTE OF MATHEMATICAL SCIENCES
FACULTY OF SCIENCE
UNIVERSITI MALAYA
KUALA LUMPUR

2024

UNIVERSITI MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: **ANEESHA PILLAY A/P BALACHANDRAN PILLAY**

Matric No: **17043604/3**

Name of Degree: **DOCTOR OF PHILOSOPHY**

Title of this Thesis ("this work"):

ON SOME METHODS OF FEATURE ENGINEERING USEFUL FOR CRANIODENTAL MORPHOMETRICS OF RATS, SHREWS AND KANGAROOS

Field of Study:

MATHEMATICS AND STATISTICS

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know thatthe making of this work constitutes an infringement of any copyright Work;
(5) I hereby assign all and every right in the copyright to this Work to the Universiti Malaya ("UM"), who henceforth shall be the owner of thecopyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date: 1/9/2024

Subscribed and solemnly declared before,

Witness's Signature Date: 1/9/2024

Name:

Designation:

ON SOME METHODS OF FEATURE ENGINEERING USEFUL FOR CRANIODENTAL MORPHOMETRICS OF RATS, SHREWS AND KANGAROOS

ABSTRACT

This study examines the craniodental morphology of biological organisms using functional data analysis (FDA). Traditional morphometrics (TM) often uses large numbers of morphometric features to study shape variation among biological organisms. However, this can lead to data redundancy, meaning that the features may contain overlapping information. This study proposes using recursive feature elimination (RFE) method to reduce data dimensionality and select the most important attributes based on predictor importance ranking. RFE was applied to the craniodental measurements of *Rattus rattus* (*R.rattus*) data to select the best feature subset for both male and female rats. A comparative study based on machine learning algorithms was also conducted by using all features and the RFE-selected features to classify the *R. rattus* sample based on the age groups. The results showed that the RFE-selected features were able to improve the classification accuracy of the machine learning algorithms. However, the linear measurements used in TM can only detect changes in size and can be insensitive to geometrical transformations. Therefore, GM is used in the subsequent work as it is more sensitive to changes in shapes. Functional data geometric morphometrics (FDGM) for 2D landmark data is introduced and its performance is compared with the classical GM method. FDGM was applied to 2D craniodental landmark data obtained from 90 crania specimens of three shrew species based on three craniodental views (dorsal, jaw, and lateral). The discrete landmarks were converted into continuous curves and represented as linear combinations of basis functions. Principal component analysis (PCA) and linear discriminant analysis (LDA) were then applied to the GM method and FDGM method to observe the classification of the shrew species. The results showed that the FDGM

approach produced better results in separating the three clusters of shrew species compared to the GM method. Machine learning approaches were also performed using predicted PC scores obtained from both methods (combination of all three craniodental views and individual views). These analyses favoured FDGM, and the dorsal view of the shrew skull was revealed to give the best representation for distinguishing between the three shrew species. This work also introduces FDGM for 3D landmark coordinate data. FDGM and GM were applied to distinguish dietary categories of kangaroos (fungivores, mixed feeders, browser, and grazer) using landmarks obtained from crania of 41 kangaroo extant species. The results showed that FDGM was able to improve the reconstruction error and distinguish dietary categories of kangaroos better than GM. Simulation studies were conducted to show the general effectiveness of FDGM compared to GM method for both 2D and 3D landmark data. The results obtained from the simulation studies and application to real data showed that FDGM performed better than GM when PCA and LDA were employed. Thus, FDGM provides a powerful and flexible framework for analysing shape variation in geometric morphometrics research.

Keywords: recursive feature elimination, traditional morphometrics, functional data geometric morphometric, principal component analysis, linear discriminant analysis.

KAEDAH-KAEDAH KEJURUTERAAN CIRI YANG BERGUNA UNTUK MORFOMETRIK KRANIODENTAL TIKUS, CENCURUT DAN KANGGARU

ABSTRAK

Kajian ini mengkaji morfologi kraniodental organisma biologi menggunakan analisis data berfungsi (FDA). Morfometrik tradisional (TM) sering menggunakan sejumlah besar ciri morfometrik untuk mengkaji variasi bentuk di kalangan organisma biologi. Walau bagaimanapun, ini boleh menyebabkan lebihan data, bermakna ciri mungkin mengandungi maklumat bertindih. Kajian ini mencadangkan penggunaan kaedah penghapusan ciri rekursif (RFE) untuk mengurangkan dimensi data dan memilih atribut yang paling penting berdasarkan kedudukan kepentingan peramal. RFE telah digunakan pada pengukuran craniodental data *Rattus rattus* (*R. Rattus*) untuk memilih subset ciri terbaik untuk kedua-dua tikus jantan dan betina. Kajian perbandingan berdasarkan algoritma pembelajaran mesin juga telah dijalankan dengan menggunakan semua ciri dan ciri yang dipilih RFE untuk mengklasifikasikan sampel *R. rattus* berdasarkan kumpulan umur. Keputusan menunjukkan bahawa ciri yang dipilih RFE dapat meningkatkan ketepatan klasifikasi algoritma pembelajaran mesin. Walau bagaimanapun, ukuran linear yang digunakan dalam TM hanya dapat mengesan perubahan saiz dan boleh menjadi tidak sensitif kepada transformasi geometri. Oleh itu, GM digunakan dalam kerja seterusnya kerana ia lebih sensitif kepada perubahan bentuk. Morfometrik geometri data fungsional (FDGM) untuk data mercu tanda 2D diperkenalkan dan prestasinya dibandingkan dengan kaedah GM klasik. FDGM telah digunakan pada data mercu tanda kraniodental 2D yang diperoleh daripada 90 spesimen krania bagi tiga spesies cencurut berdasarkan tiga pandangan kraniodental (dorsal, rahang dan sisi). Tanda tempat diskret telah ditukar kepada lengkung berterusan dan diwakili sebagai gabungan linear fungsi asas. Analisis komponen utama (PCA) dan analisis diskriminasi linear (LDA) kemudiannya digunakan pada kaedah GM dan kaedah FDGM untuk memerhati

klasifikasi spesies cencurut. Keputusan menunjukkan bahawa pendekatan FDGM menghasilkan keputusan yang lebih baik dalam mengasingkan tiga kelompok spesies cencurut berbanding kaedah GM. Pendekatan pembelajaran mesin juga dilakukan menggunakan skor PC ramalan yang diperoleh daripada kedua-dua kaedah (gabungan ketiga-tiga pandangan kraniodental dan pandangan individu). Analisis ini mengutamakan FDGM, dan pandangan dorsal tengkorak cencurut telah didedahkan untuk memberikan perwakilan terbaik untuk membezakan antara tiga spesies cencurut. Kerja ini juga memperkenalkan FDGM untuk data koordinat mercu tanda 3D. FDGM dan GM telah digunakan untuk membezakan kategori diet kanggaru (fungivor, penyuap campuran, memakan lebih daun dan batang dikotil, dan hanya memakan lebih rumput) menggunakan tanda tempat yang diperoleh daripada crania 41 spesies kanggaru yang masih wujud. Keputusan menunjukkan bahawa FDGM dapat memperbaiki ralat pembinaan semula dan membezakan kategori pemakanan kanggaru lebih baik daripada GM. Kajian simulasi telah dijalankan untuk menunjukkan keberkesanan umum FDGM berbanding kaedah GM untuk kedua-dua data mercu tanda 2D dan 3D. Keputusan yang diperoleh daripada kajian simulasi dan aplikasi kepada data sebenar menunjukkan bahawa FDGM menunjukkan prestasi yang lebih baik daripada GM apabila PCA dan LDA digunakan. Oleh itu, FDGM menyediakan rangka kerja yang berkuasa dan fleksibel untuk menganalisis variasi bentuk dalam penyelidikan morfometrik geometri.

Kata kunci: penghapusan ciri rekursif, morfometrik tradisional, morfometrik geometrik data berfungsi, analisis komponen utama, analisis diskriminasi linea

ACKNOWLEDGEMENTS

First and foremost, I would like to express my heartfelt gratitude to my supervisors, Dr. Dharini Pathmanathan, Professor Dr. Sophie Dabo- Niang, Dr Arpah binti Abu and Associate Professor Dr Hasmahzaiti binti Omar. Their invaluable guidance, assistance and for their direction, invaluable guidance assistance and unwavering support have been instrumental in shaping my academic journey and enabling me to achieve this milestone. Their expertise and dedication to my research have been truly inspiring, and I am forever grateful for their mentorship.

I would also like to extend my sincere appreciation to Nurul Aityqah binti Yaacob from Universiti Malaya for her invaluable advice and suggestions throughout my thesis completion. Her insights and support have been crucial in refining my research and propelling me towards success. I am also immensely grateful to Shafiqah Azman and Khoo Tzung Hsuen, for their invaluable contributions to my research.

This thesis is dedicated to my late father, D.H. Balachandran Pillay, whose unwavering belief in my abilities has been a constant source of motivation. His absence has been deeply felt, but his legacy continues to inspire me to pursue my dreams with determination. I wish to record my deepest gratitude to my mother, Subathra Nair, whose unwavering love, prayers and encouragement have been my pillars of strength. I am immensely grateful to my brother, Anesh Pillay for his constant encouragement, which has been a source of strength and inspiration. Thank you for all the love, encouragement, and financial support throughout this study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

I_x	:	Compact set in $\mathbb{R}$, with finite Lebesgue-measure
$C(\mathrm{s},\mathrm{t})$	:	Covariance function
Q	:	Covariance matrix of the variables centered on the class mean
Γ	:	Covariance operator
ϕ_j	:	Eigenfunctions of covariance operator
λ_j	:	Eigenvalues of covariance operator
a	:	Eigenvector of B corresponding to the largest eigenvalue
$\hat{\phi}_{p,j}(t_p)$	:	Elements of the estimated multivariate eigenfunctions
$\hat{\phi}_{p,j}$	:	Estimated eigenfunctions
ρ	:	Selected principal component scores
$\mathcal{H}$	:	Hilbert space
G	:	Matrix of class indicator variables
$\bar{\rho}$	:	Mean of the principal component scores over the whole sample
X	:	Multivariate functional data
ξ	:	Multivariate functional principal component scores
$\mathrm{X}(\mathrm{t})$	:	Multivariate version of the Karhunen-Loève's representation
P	:	Number of domains
N	:	Number of landmarks
n	:	Number of specimens
τ	:	Observed landmark points
J	:	Orthonormal basis function
$\hat{\mathrm{v}}_\mathrm{j}$	:	Orthonormal eigenvectors
%	:	Percentages

B	:	Predictions by class means
$\mathbb{R}$	:	Real numbers
W	:	Sample covariance matrix
$\langle\langle\cdot,\cdot\rangle\rangle$	:	Scalar product
M	:	Set of classes
$\mathcal{L}^2(I_x)$	:	Space of square integrable functions on I_x
$\mathrm{X}_{[J]}(\mathrm{t})$	:	Truncated Karhunen-Loève expansion of process **X**
$X_{p,[J_p]}(t_p)$	:	Truncated Karhunen-Loève expansion of the components of **X**
X_p	:	Univariate function
$X(\cdot)$	:	Univariate functional data sample for $x-$coordinates
$Y(\cdot)$	:	Univariate functional data sample for $y-$coordinates
$Z(\cdot)$	:	Univariate functional data sample for $z-$coordinates
$K_p(\cdot,\cdot)$	:	Variance-covariance function
2D	:	Two dimensional
3D	:	Three dimensional
Acc	:	Accuracy
AUC	:	Area under the receiver operating characteristic curve
ANN	:	Artificial neural network
BBP	:	Breadth across palate at first molar
BBC	:	Breadth of braincase
BM1	:	Breadth of first upper molar
BIF	:	Breadth of incisive foramina
BMF	:	Breadth of mesopterygoid fossa

BR	:	Breadth of rostrum
BZP	:	Breadth of zygomatic plate
CLM1.3	:	Crown length of maxillary molar row
FN	:	False negative
FP	:	False positive
FDA	:	Functional data analysis
FDGM	:	Functional data geometric morphometric
FLDA	:	Functional linear discriminant analysis
GLM	:	Generalised linear model
GM	:	Geometric morphometric
HBC	:	Height of braincase
IB	:	Interorbital breadth
LB	:	Length of auditory bulla
LBP	:	Length of bony plate
LD	:	Length of diastema
ML	:	Length of mandible
M1.M3	:	Length of mandible toothrow
LR	:	Length of rostrum
LDA	:	Linear discriminant analysis
MFPCA	:	Multivariate functional principal component analysis
NB	:	Naïve Bayes

ONL	:	Occipitonasal length
PPL	:	Post palatal length
PCA	:	Principal component analysis
RF	:	Random forest
ROC	:	Receiver operating characteristic curve
RFE	:	Recursive feature elimination
SVM	:	Support vector machine
TIFF	:	Tagged Image File Format
TM	:	Traditional morphometric
TN	:	True negative
TP	:	True positive
UM	:	Universiti Malaya
ZB	:	Zygomatic breadth

LIST OF APPENDICES

CHAPTER 1: INTRODUCTION

1.1 Background

Morphometrics is a fundamental discipline in biological research that focuses on quantitatively describing and analysing shape and its variations across organisms (Rohlf, 1990). Initially centered on basic descriptive measurements, this field has progressed significantly and is currently employing advanced statistical and computational techniques to study shape and size variation (Adams et al., 2013). The importance of morphometrics transcends disciplinary boundaries, finding applications across various biological domains such as evolutionary biology, ecology, anthropology, biomedical sciences, and other fields, underscoring their versatility and utility (Slice, 2005). In ecology and evolutionary biology, morphometric analyses have provided insights into the processes underlying phenotypic diversification, speciation, and adaptation (Adams & Otárola-Castillo, 2013). In taxonomy and systematics, morphometric approaches facilitate species delimitation and phylogenetic reconstructions, enhancing understanding of the evolutionary relationships among organisms (Swiderski et al., 2004). Moreover, in biomedical sciences, morphometrics plays a vital role in medical imaging, diagnostics, and treatment planning, aiding in the understanding and management of various health conditions (Bookstein, 1996).

Conceptually, morphometrics can be broadly categorised into three approaches: traditional morphometrics (TM) that relies on linear distance measurements of biological organisms for statistical analysis, landmark-based morphometrics that requires precise positioning of anatomical landmarks, and outline-based morphometrics which captures the contour of forms through a sequence of pseudo-landmarks (Dujardin, 2017; Rohlf & Marcus, 1993). As morphometric techniques continue to expand, the selection of appropriate methods becomes crucial for meaningful applications in biological research.

Traditional morphometrics is a foundational method used in the study of biological shape variation. It involves mathematical concepts and geometric reasoning to explain a wide range of biological phenomena, providing insights into the processes underlying morphogenesis and evolution (Thomson, 1917). This approach applies multivariate statistics to sets of morphological variables such as linear distances between landmarks and sometimes angles, ratios etc. The TM approach was followed by an era where the study on coordinates of landmarks and the geometric information about their relative positions led to the innovation in morphometrics through the introduction of the geometric morphometrics (GM) method.

Geometric morphometrics, a technique developed by (Bookstein, 1984, 1986, 1987, 1991) is a popular method for studying morphological variation in biological organisms. Unlike TM, which relies on linear measurements, GM is based on the idea that the shape of an organism can be described by the coordinates of a set of landmarks on its surface. Landmarks are points on the image of the organism that are consistently located in the same place, regardless of the size or orientation of the organism (Slice, 2005). Landmarks are categorised into three types, defined by biology (Type I), geometry (Type II), and relative positions (Type III) (Bookstein, 1991) although Bookstein later redefined Type III landmarks as semi landmarks (Bookstein, 1997). Type I landmarks are points located at anatomically homologous locations across specimens. These landmarks are easily identifiable and show little variation in position across individuals within a species. Type II landmarks are points that may not be homologous across specimens but are meaningful for describing shape variation. These landmarks are often placed along curves or outlines of structures. Type III landmarks are semilandmarks placed along curves or outlines where there are no clear anatomical points. These landmarks are typically evenly spaced along curves or outlines and are used to capture overall shape variation. Different types

of landmarks are chosen based on the characteristics of the biological structures being studied, and the level of detail required to capture shape variation effectively.

The concept of Procrustes superimposition as a fundamental technique in GM for analysing shape variation, involves aligning landmark configurations by removing differences in translation, rotation, and scaling to enable direct comparison of shape (Rohlf & Marcus, 1993; Slice, 2005). Generalised Procrustes analysis (GPA) is be applied on raw landmarks to superimpose the landmark configurations using least-squares estimates and rotation parameters (Adams et al., 2004). These variables can be used to compare the shapes of different organisms using graphical visualisation of results to track changes in shape over time and to identify the underlying causes of shape variation.

The shift from GM to outline morphometrics (OM) represents an evolution in the methods used to capture and analyse shape variation in biological structures. While both approaches focus on quantifying shape differences, they differ in the way shapes are represented and analysed. Geometric morphometrics primarily relies on the identification and digitisation of anatomical landmarks on biological structures. On the other hand, OM focuses on capturing the overall shape of an object based on a series of pseudo-landmarks that describe contours or boundary outlines without depending on the presence of true anatomical landmarks (Dujardin et al., 2014). Elliptical Fourier analysis (EFA), developed by (Kuhl & Giardina, 1982) is one of the established methods of OM that is particularly useful for analysing shapes with smooth, continuous outlines. This mathematical tool decomposes the outline of a shape into a series of sine and cosine curves using Fourier transforms (Caple et al., 2017) which capture the variation in shape along the outline, allowing for the quantification and comparison of shape differences. Another common approach in OM is the thin-plate splines (TPS) (Bookstein, 1987). This technique interpolates and warps one shape into another by minimising bending energy, allowing for the visualisation and quantification of shape changes between outlines.

Despite its broad utility, morphometrics presents several methodological challenges and considerations. These include issues related to data acquisition, such as ensuring the accuracy and reproducibility of measurements, as well as statistical analyses, such as dealing with high-dimensional data and controlling for potential sources of bias and error. Furthermore, the interpretation of morphometric results can be complex, requiring careful consideration of biological context and ecological factors.

Feature engineering is essential in morphometrics studies to select informative variables or features from raw data that capture relevant aspects of shape variation in biological organisms. Morphometric data often exhibit high dimensionality, multicollinearity, and noise, which can pose challenges for analysis. Therefore, feature engineering helps researchers determine interpretable features to understand the morphological differences between groups, identify key factors influencing shape variation, and generate hypotheses about evolutionary, developmental, or ecological processes. Furthermore, feature engineering techniques such as dimensionality reduction, feature selection, and transformation can help to reduce noise, redundancy, and overfitting in morphometric models.

In TM, researchers often measure numerous linear distances and angles. Therefore, feature engineering comes into handy in selecting the most informative variables while discarding redundant or irrelevant ones, thus effectively reducing the dimensionality of the data. Recursive feature elimination (RFE) is a feature selection technique that iteratively removes less important variables from the dataset until the optimal subset of features is identified. This technique applies a backward selection process that starts with the full set of features and iteratively removes the least important features in a data set. RFE trains a model iteratively, ranking the features according to their importance scores and then removing the lowest ranking predictors (Darst et al., 2018). The application of RFE is incorporated in my thesis to observe its efficiency to determine the best feature

subset using the craniodental linear measurements in TM.

This thesis underscores the significance of craniodental morphology, which encompasses the study of the skull (cranium) and teeth (dental) shape and structure in vertebrates, particularly rats, shrews, and kangaroos. Craniodental morphology is pivotal for elucidating evolutionary relationships, ecological adaptations, and functional aspects across species. Additionally, craniodental morphology serves as a framework for modeling morphological evolution in both modern and fossil lineages within phylogenetic analyses (Cardini & Elton, 2008). Insights gleaned from the shape and size of craniodental structures offer valuable information regarding adaptations to specific ecological niches and specialised feeding behaviors (Tse & Calede, 2021). This thesis endeavors to apply feature engineering techniques for TM analysis on craniodental linear measurements of rats, while also proposing an alternate GM approach based on functional data analysis (FDA) to investigate craniodental structures of shrews and kangaroos.

In my thesis, FDA is incorporated in GM to observe classification accuracy among biological organisms. FDA is a statistical methodology utilised to analyse data represented in the form of functions, such as curves or surfaces, rather than discrete observations. It is particularly advantageous for handling data that exhibit continuous variation over a domain, such as time, space, or wavelength. In this thesis, FDA is employed to represent discrete observations, such as landmark coordinates, as functions. This transformation involves creating functional data that encapsulates all the coordinates as a single observation, thereby capturing the entire measured function. Subsequently, models are developed to predict information based on a collection of functional data, utilising statistical principles from multivariate data analysis (Ullah & Finch, 2013).

Functional data geometric morphometrics (FDGM) is proposed in this thesis, requiring steps to perform statistical analysis on signals, curves, or even more complex objects

while being invariant to certain shape-preserving transformations (Gu et al., 2022). To address the need for alignment of functions in geometric features like peaks and valleys, curve registration (Ramsay & Li, 1998; Srivastava et al., 2011) or functional alignment (Ramsay, 2006) techniques are applied. These methods warp the temporal domain of functions to ensure proper alignment, enabling accurate analysis of geometric features (Guo et al., 2022). The proposed method involves the development, implementation, and verification of FDGM which includes a set of statistical models' alternative to multivariate models by transforming large complex data into functional data such as data objects, curves, shapes, images, or a more complex mathematical object. The statistical goals of this study then include comparisons, summarisation, clustering, modeling, and testing of functional and shape skulls objects.

In addition, this study also incorporates machine learning into morphometric studies for taxonomic classification. Commonly used classification methods include naïve Bayes (NB), random forest (RF), generalised linear model (GLM), support vector machine (SVM) and artificial neural network (ANN).

The methods involve collecting data which includes either the linear craniodental measurements directly obtained from skulls or 2D and 3D landmark data from the skull images of biological organisms. TM and GM studies will be performed and the FDGM approach will be implemented and tested. The FDGM approach is developed, and it aims to bring a real added value to the problem of interest.

1.2 Problem Statements

The study of the shapes of biological organisms is a challenging task, as the shapes are often complex and difficult to quantify. Quantitative approaches allowed scientists to study shapes of various organisms better where they no longer rely on word descriptions which lead to different interpretations. TM method has been widely used in identification of species, analysis of morphological characters and other parts of taxonomy. Traditionally, variables used in morphometric analysis are linear distances between landmarks which are directly measured on the specimens. This method also used angles, counts, ratios and areas. However, TM can be difficult to capture the full geometry of an object using linear measurements. For example, the shape of a skull is determined by the size, shape, and orientation of the bones that make up the skull. The distances between landmarks on the skull can only capture some of this information. In addition, the distances between landmarks can be affected by the size of the object. For example, it is difficult to compare the shapes of two skulls if one skull is twice as large as the other skull. The data may also contain less information due to directions measured redundantly and most of these measurements tend to overlap. GM overcomes these limitations by using coordinate-based data to capture the shape of an object. This allows for a more comprehensive description of the shape of an object.

The shift from GM to OM reflects a recognition of the limitations of landmark-based methods in capturing certain types of shape variation, particularly in structures with complex or continuous outlines. Outline morphometrics offers a more flexible approach to shape analysis, allowing researchers to quantify shape variation in a wider range of biological structures.

Additionally, outline morphometrics can complement geometric morphometrics by providing a more comprehensive analysis of shape variation, especially in cases where landmark-based methods may not fully capture the nuances of shape differences. By incorporating both landmark-based and outline-based approaches, researchers can gain a more complete understanding of shape variation in biological structures and address a broader range of research questions.

It is of my interest to explore morphometrics of craniodental characters based on the functional data analysis (FDA) approach. FDA is a branch of statistics that analyses data that is naturally ordered or structured. This type of data is often encountered in morphometrics, where the shapes of objects are represented as curves or surfaces. The main advantage of FDGM over GM is that it can be used to analyse data that varies over time or space. For example, FDGM can be used to study the changes in the shape of a skull over time or the differences in the shape of skulls between different populations also be a more general statistical approach than GM. This implies that it can be used to analyse a wider variety of data types and to answer a wider variety of research questions which makes FDA a more powerful and versatile than GM. This study aims to develop and implement a functional data geometric morphometric (FDGM) approach to study the skull shapes of biological organisms.

The FDGM approach will be compared to other morphometrics methods. This project focuses on incorporating FDA in the form of functions for shape analysis based on 2D and 3D landmark data. Simulation studies for both 2D and 3D landmark data were also conducted to test the general effectiveness of the FDGM approach compared to GM.

1.3 Significance of Research

Due to the presence of redundant linear measurements in the TM approach, a good feature selection method should be used in the study to select the best, highly discriminant features, which can increase the performance of the model and reduces computational complexity in classification problems. This study revealed that RFE-based feature selection technique can classify biological organisms better when incorporated in the TM approach. RFE has proven to be advantageous the most relevant features in predicting the target variables, thus this study hypothesises that this method would also benefit in TM studies to classify among groups among biological organisms.

Besides that, this study also proposes a new and more exhaustive way to see, manipulate and study the skulls of biological organisms where a data (unit) is not a vector (multivariate), but all available information including its dynamics (shape). There is also a limited number of FDA models available for the explanation, visualisation, classification, and modelling of the geometric morphometric dataset of interest. The major statistical challenge is to pay attention to hot topics such as the management of missing or low quality of data (e.g., a part of the shape).

In addition, the implementation of an FDA approach requires a correct definition of the targeted function spaces, appropriate metrics to measure the similarity between objects, spatial correlation techniques etc. Indeed, one of the main challenges in the analysis (dimensional reduction, regression models, tests) of large complex data is to use statistical tools capable of performing calculations in an inexpensive way with correlations among huge amounts of data (Chen et al., 2011; Zipunnikov et al., 2011).

1.4 Aims and objectives

The general aim of this study is to introduce an FDA approach in morphometric studies to analyse craniodental characters.

The objectives of this morphometric study are to:

i. Incorporate and review random forest recursive feature elimination as a feature engineering method into traditional morphometrics to overcome data redundancy.

ii. propose FDA-based framework as a feature engineering technique to enhance geometric morphometrics for 2D and 3D skull shape analysis in detecting variation among biological organisms.

iii. conduct comparative studies using machine learning on FDA- based 2D and 3D FDGM to discriminate between groups of biological organisms.

1.5 Research Outline

This research re-evaluates the TM method by incorporating RFE-based feature selection technique to observe the classification accuracy of selected linear measurements. This study also introduces the application of FDA in GM and proposes FDGM that incorporates this approach into 2D and 3D landmark data in GM. The research is outlined as follows:

Chapter 2 provides a literature review of the TM, GM and FDA approaches used in previous studies. Then, the machine learning algorithms were also reviewed.

Chapter 3 addresses the application of the RFE as a feature selection technique in TM. RFE was applied to observe age classification among male and female *R.rattus* rats in Peninsular Malaysia. A comparison study was conducted to examine the effectiveness of RFE-selected features with all linear measurements obtained using machine learning algorithms.

Chapter 4 proposes the FDGM method into 2D geometric morphometric. FDA approach is incorporated into 2D craniodental landmark data of three shrew *species* (*C. malayana, C. monticola* and *S. murinus*). Machine learning algorithms and simulation studies were also used to assess the accuracy of the proposed approach.

Chapter 5 describes the extension of the FDGM method into 3D geometric morphometrics. Using a train-test ratio of 70:30, the effectiveness of the proposed method is examined using machine learning algorithms and simulation studies.

Chapter 6 provides concluding remarks and some significant contributions from this research. Suggestions on extending research work related to this research are also included in this chapter.

CHAPTER 2: LITERATURE REVIEW

2.1 Morphometrics

The study of form plays a crucial role in biological research. Morphometrics is the statistical study of shape variation and covariation with other variables (Bookstein, 1996; Dryden & Mardia, 1998). During the 1960s and 1970s, biometricians employed multivariate statistical methods to explore the realm of morphometrics (e.g., Giles & Elliot, 1963; Birkby, 1966; Rohlf, 1972; Van Valen, 1974; Albrecht, 1979). Blackith and Reyment (1971) were instrumental in delineating a spectrum of multivariate statistical techniques tailored for the domains of biology and paleontology. Statistical methodologies such as discriminant functions, canonical variates, principal components analysis (PCA), factor analysis, cluster analysis, and trend surface analysis were also discussed in this work, thereby primarily furnishing biologists with a foundational framework to adopt multivariate methods in their research (Blackith & Reyment, 1971). Biology underwent a transformation from descriptive to quantitative approaches, and the field of morphology mirrored this quantification revolution (Bookstein, 1998). In paleontology, morphological differences and distances serve as the primary criteria for distinguishing between species and genera (Stafford & Szalay, 2000). Statistical methods such as the correlation coefficient, analysis of variance and principal component analysis further advanced quantitative rigour. The sophistication of these analyses evolved in tandem with the rapid advancements in statistics.

2.2 Traditional Morphometrics

Traditional morphometrics (TM) plays a pivotal role in understanding morphological variation among biological organisms through the meticulous analysis of linear distances, angles, ratios, and other morphological variables.

These measurements, although often correlated with size, serve as fundamental descriptors of shape, once size effects are mitigated. Conventional morphometrics using linear distance measurements of skulls have proven to be powerful for identification, classification, and analysis of skull variability among biological organisms. Many researchers conducted TM using linear measurements which are directly obtained from biological organisms. These measurements are later analysed using multivariate statistical approaches to identify the morphological variation among groups of individuals (Chuanromanee et.al, 2019). For instance, Howells (1989) employed PCA to scrutinise metric dental variation across major human populations, shedding light on population-level distinctions.

Brace and Hunt (1990) employed C scores, derived from craniofacial measurements across diverse populations from Asia, the Pacific, the aboriginal western hemisphere, and Europe. Their work, which culminated in Euclidean distance dendrograms, revealed distinct regional clusters, offering insights into the degrees of relationship among populations based on nonadaptive traits (Brace & Hunt, 1990). The methodological richness of TM is further underscored by Marcus (1990), who provided a comprehensive overview of analytical techniques ranging from PCA to discriminant analysis. This work not only elucidated the application of these methods but also addressed crucial aspects such as resampling techniques for robust estimation of standard errors (Marcus, 1990). Moreover, TM is not confined to anthropological studies alone; it transcends disciplinary boundaries. Abdelhady and Elewa (2010), for instance, utilised TM to study the evolution of *Exogyrinae* oysters. Through PCA, cluster analysis, and cladistic analysis, they uncovered morphological dimorphism between species and delineated temporal boundaries for the examined oyster members (Abdelhady & Elewa, 2010).

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., 2016). 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 generalisation models (Li et al., 2016).

2.3 Geometric Morphometrics

Geometric morphometrics has emerged as a powerful tool for analysing and quantifying biological shapes, gaining widespread popularity for studying morphological variation in diverse organisms, including fish, birds, mammals, and insects. Its introduction in the 1980s revolutionised morphometrics, shifting the focus from traditional measurements to landmark-based geometric information. In GM, landmarks are pivotal for analysing and quantifying the shapes of biological structures. They serve as reference points that allow for the comparison of shapes across different specimens. Researchers identify specific anatomical points on each specimen that correspond to the defined landmark types (Type I, Type II, and Type III). These landmarks are chosen based on their biological significance and their ability to be consistently located across all specimens being studied. Landmarks (Type I) are discrete and anatomically homologous points that can be precisely located on every specimen. They are defined by clear anatomical features such as intersections of sutures or the tips of structures. These landmarks can be consistently identified across different specimens and observers. Type I landmarks provide the most accurate points for aligning shapes because of their clear and consistent anatomical basis. Pseudo landmarks (Type II) are points located on geometric features such as the maxima of curvature or along the boundary of a structure. They are not as precisely defined anatomically as Type I landmarks but still provide

important geometric information. Type II landmarks are useful for describing the general shape and form of structures. Semilandmarks (Type III) are points that are placed along curves or surfaces where precise homologous points are difficult to identify. They are defined relative to other landmarks or along a structure. These points are particularly useful for capturing the shape of curves and surfaces where precise homologous points cannot be identified. They allow for more flexible and comprehensive shape analysis, especially for complex structures. Using specialised software, or imaging techniques, the coordinates of these landmarks are recorded. This process converts the physical shape of the specimens into a numerical format that can be analysed mathematically. There are many open-source and licensed software that are available for GM analysis. Table 2.1 includes commonly used software for landmark digitising and GM.

Table 2.1: Available software for geometric morphometric analysis.

Types	**Software**	**Sources**
Landmark digitising	tpsDig2 TINA Manual Landmarking Tool 3Skull	Rohlf (2017) Schunke et al. (2012) Ousley (2004)
Geometric morphometrics	MorphoJ R package, geomorph	Klingenberg (2011) Adams & Otarola-Castillo, 2013; Adams et al. (2018)

Bookstein (1984) presented a pioneering landmark statistical approach that outlines the theoretical framework behind tensor method, demonstrating how it can be applied to analyse shape differences among various biological entities. (Bookstein, 1984). This approach enables researchers to focus on shape variations independent of size or location, facilitating meaningful comparisons across samples (Bookstein, 1984). Landmarks are aligned using techniques such as Procrustes superimposition. This involves translating, rotating, and scaling the landmark configurations to a common coordinate system.

The goal is to minimise differences that are not related to shape (e.g., size, orientation) and to focus solely on shape differences.

Kendall (1984) laid the groundwork by demonstrating that when the vertices of a shape adhere to independent and identically distributed spherical normal variables, the resulting distribution of the shape becomes isotropic across Kendall's shape space. This work safeguards against the distortion of shape space by isotropic measurement errors, ensuring the integrity of analyses (Bookstein, 1991; Dryden & Mardia, 1998; Mitteroecker & Gunz, 2009). Two-point registration, also known as Bookstein's shape coordinates, is a straightforward superimposition method that significantly influenced Bookstein's development of shape theory in the late 1980s. Generalised Procrustes analysis (GPA) aligns landmark configurations using least-squares estimates for translation and rotation parameters. This process begins by translating the centroid of each configuration to the origin, followed by scaling the configurations to a common unit size by dividing by the centroid size (Adams et al., 2004; Bookstein, 1986). Finally, the configurations are optimally rotated to minimise the squared differences between corresponding landmarks (Adams et al., 2004; Gower, 1975; Rohlf & Slice, 1990). This process is repeated iteratively to compute the mean shape, which cannot be estimated before superimposition. After superimposition, shape differences can be described by the differences in the coordinates of corresponding landmarks between objects. These differences can also be utilised as data in multivariate comparisons of shape variation.

A significant portion of the foundational work in GM was published between 1981 and 1991 (Macleod, 2017). Bookstein (1991) provides a comprehensive introduction to geometric morphometrics, covering the mathematical foundations and biological applications of landmark-based analysis, which serves as a foundational reference for researchers entering the field. Moreover, this work underscores the importance of linking geometric patterns to evolutionary processes, ecological interactions, and developmental

mechanisms, thereby enriching the understanding of shape variation in biological systems (Bookstein, 1991).

Rohlf and Marcus (1993) highlighted the transformative impact of geometric morphometrics by elucidating how landmark-based methods, such as superimposition methods that emphasises applications to exploratory studies in taxonomy and evolution. Their work encompassed a thorough examination of various procedures utilised in describing shape in biology and have termed the use of GM as a revolution in describing the "shape" (Rohlf & Marcus, 1993). Adams et al. (2004) revisited the foundational principles of GM introduced in the earlier 'revolutionary' work and highlighted key developments in methodology, theory, and applications of the approach. Notably, their study underscored advancements in landmark selection, superimposition techniques, shape visualisation, and statistical modeling, illustrating the continuous evolution of GM techniques. Additionally, Adams et al. (2004) explored the synergistic integration of GM with other quantitative approaches, such as phylogenetic comparative methods and quantitative genetics, emphasising the interdisciplinary nature of morphometric research and its potential for enriching biological inquiries fostering interdisciplinary collaborations and enriching the scope of morphometric research.

Webster and Sheets (2010) introduced common exploratory and confirmatory techniques in landmark-based geometric morphometrics. This paper also covers issues that are frequently faced by biologists in comparative morphology studies and focuses in 2D and 3D landmark data analysis. Besides that, it also covers the topics of acquiring landmark data, superimposition methods, visualising shape variation, quantifying and statistically comparing the amount of shape variation, statistical testing of difference in mean shape, and statistical assignment of specimens to groups.

Within morphometrics, craniodental morphology holds particular significance, offering insights into taxonomic discrimination, evolutionary studies, and biomedical implications. Adams and Rohlf (2000) highlighted the importance of craniodental morphometrics in elucidating ecological character displacements in *Plethodon* salamanders through landmark-based geometric morphometric analysis (GM). Slice (2005) explored the application of morphometrics in physical anthropology with a significant focus in craniodental morphonology. The work highlighted the use of landmark-based morphometrics in studying human evolution and practical application in anthropology. These studies not only shed light on the functional adaptations of craniodental structures but also serve as inspiration for further extending the GM technique for craniodental morphology of this paper.

The efficiency of GM shines through in numerous studies. Maderbacher et al. (2008) showcased the superior efficiency of geometric morphometrics GM compared to TM in discriminating between populations of *Tropheus moorii*. Their research highlighted the limitations of TM, including its lack of diagnostic power and time-consuming nature, while emphasising GM's flexibility in terms of data acquisition and robustness as an alternative approach. Furthermore, Maderbacher et al. (2008) demonstrated that canonical variate analysis using GM data, particularly incorporating semi-landmarks, offered the most informative description of morphological differences among populations. Arias-Martorell, et al. (2015) analysed the shape of the shoulder joint (proximal humerus and glenoid cavity of the scapula) of three australopith specimen using 3D geometric morphometrics. Marcy et al. (2015) also captured 19 crania of Australia's smallest rodent using 3D scanner and μCT scanner for geometric morphometrics to classify the specimens based on sexual dimorphism. Dudzik (2019) also used GM to examine the cranial morphology of Asian and Hispanic populations by performing discriminant and canonical variate analyses.

The results of the GM analysis revealed significant differences in cranial shapes between the two groups, yet both studies concur that GM serves as a valuable tool for identifying morphological similarities among populations based on cranial morphology (Dudzik, 2019).

In another review work, Adams & Otárola-Castillo (2013) highlighted the development of morphometrics related to the Procrustes paradigm and the methodological toolkit of geometric morphometrics. The use of three-dimensional data in geometric morphometric gained a lot of popularity where there are no mathematical limitations for handling data but algorithms for superimposition, projection, and statistical analysis are all generalised to accommodate data of any dimensionality (Adams & Otárola-Castillo, 2013). Initially three-dimensional data required the use of expensive equipment and the use of devices related to it were limited. However, low-cost options such as surface scanners and other devices have become available (Adams et al., 2013). Since then geometric morphometrics using three-dimensional data became more popular.

Mitteroecker and Schaefer (2022) reviewed the recent developments and current methodological challenges of GM for biological meaningfulness. Promising directions for further research and evaluation of new developments were also outlined and illustrated on 3D human face shape based on data obtained from Avon Longitudinal Study of Parents and Children (ALSPCA) (Mitteroecker & Schaefer, 2022). Zhang et al. (2023) successfully applied GM using 2D landmarks to distinguish two subgenera classification of *Chaetocnema*, which should that GM could be used to detect morphological delimitation of the supraspecies taxa.

While GM offers powerful tools for quantifying shape variation, it is not without limitations. One critical drawback is its sensitivity to landmark placement and digitisation errors, which can introduce variability and compromise the accuracy of shape analyses.

Martensson (1998) addressed the challenges posed by measurement error in GM. This work also explores the sources of measurement error in GM, particularly focusing on the issues related to landmark placement and digitisation and offers empirical strategies to assess and mitigate the impact of these errors on shape analysis. In addition, a study by Robinson et al. (2002) investigated the impact of landmark placement error on shape analyses study of tooth shape using GM, by calculating its effect on the recorded variation in Procrustes fits, obtained for each set of multiple representations. They demonstrated that discrepancies in landmark positioning can lead to variation in orientation, thus affecting the outcomes of statistical analyses (Robinson et al., 2002).

Another issue lies in the application of Procrustes superimposition in GM. Sheets and Webster (2010) highlighted concerns about disregarding the orientation of biologically relevant axes during rotation in Procrustes superimposition which can lead to variations in the relative orientation of symmetrical axes within samples, thus complicating the description of shape differences in relation to the axis of symmetry. Additionally, their work also pointed out the concern of the "Pinocchio Effect" in this superimposition method, where large differences at some landmarks are spread out over many landmarks during the least-squares rotation, assuming equal variance at all landmarks. Despite these limitations, the study recommended the application of Procrustes methods in GM in studies for their statistical robustness (Rohlf, 2000; Sheets & Webster, 2010).

Furthermore, factors such as sample size and the selection of views and elements in two-dimensional geometric morphometric (2DGM) analyses pose additional challenges. Rummel et al. (2024) explored the influence of sample size on mean shape, shape variance, and the concordance of multiple skull 2D views in the study of bat species. Their findings underscored the importance of adequate sample sizes and careful selection of views and elements for accurate analyses (Rummel et al., 2024).

In response to these limitations, researchers have extended their methods to include outline-based morphometrics, offering alternative approaches to address some of the challenges associated with traditional GM techniques. These efforts reflect ongoing endeavors to improve the reliability and robustness of shape analysis methods in biological research.

2.4 Outline-based Geometric Morphometrics

Outline-based GM focuses on the analysis of shapes based on the outlines of objects or structures. This approach offers several advantages, including the ability to capture complex shapes and the potential for automation in data collection and analysis. Kuhl and Giardina (1982) outlined a direct procedure for obtaining the Fourier coefficients of a chain-encoded contour, emphasising its advantages that it does not require integration, or the use of fast Fourier transform techniques, and that bounds on the accuracy of the image contour reconstruction are easy to specify. The study also discussed the extension of contour representation to encompass arbitrary objects at diverse aspect angles (Kuhl & Giardina, 1982). These procedures are positioned as directly applicable to a range of pattern recognition challenges that entail analysing clearly defined image contours.

One of the pioneering works in outline-based GM is the study by Bookstein (1991), where key conceptual frameworks relevant to both landmark-based and outline shape analysis have been highlighted such as the use of shape coordinates and thin plate splines. Bookstein (1991) is significant for laying the groundwork for outline shape analysis by addressing the challenges of analysing shapes that are not easily defined by discrete landmarks. This work also touches on variants of the general procedure encountered in the outline processing which are taking derivatives of the outline curves and measuring dissimilarity between forms in terms of squared differences of those derivatives rather than distances between the original paired point loci. Additionally, Bookstein (1990) also discusses how information from curving outlines can be analysed effectively once the

landmarks are dealt with. MacLeod (2007) established a solid theoretical foundation for understanding the principles behind automated taxon identification. His study described the statistical and computational underpinnings necessary for developing reliable identification systems. Besides that, the study also explained the use of image analysis for species identification, by describing how digital images of specimens can be processed and analysed to extract distinguishing features that can be used for taxon identification (MacLeod, 2007). Dujardin et al. (2014) demonstrated the outline method's efficacy in distinguishing close or cryptic species and characterising conspecific geographic populations across various vector organisms. Notably, in recognising such forms, the study observed that the outline approach yields comparable results to the landmark-based method (Dujardin et al., 2014).

2.5 Functional Data Analysis

This research aims to provide an alternative to the GM multivariate approach, which is functional data analysis (FDA) that includes a set of statistical techniques considering the structured data of interest into shape objects, thought of as smooth realisations of a stochastic process (Hall & Vial, 2006; Srivastava & Klassen, 2016). FDA based on the landmark method aligns special features in functions or derivatives to their average location and then smooth to the location of the feature (Kneip & Gasser, 1992; Gasser & Kneip, 1995). Bookstein (1997) introduced a combination of Procrustes analysis and thin-plate splines, the two most powerful tools of landmark-based morphometrics, for the multivariate analysis of curving outlines in MRI images of the human brain. This work effectively describes group differences in data from curving forms that do not need to have any reliable point-like landmarks anywhere along the arcs. The method works by treating the thin-plate splines and Procrustes fitting as a nonlinear filter for regional differences in outline shape, with their bandpass characteristics complementing each other directionally. This complementary filtering enhances the effectiveness of Procrustes

analysis following spline-based preprocessing (Bookstein, 1997). Ramsay and Silverman (2005) provided a comprehensive introduction to FDA, covering theoretical foundations and practical applications, including methods for clustering and classification of functional data, which is particularly relevant for grouping similar shapes or curves in morphometrics (Ramsay & Silverman, 2005). The FDA framework allows better accuracy in parameter estimation in the analysis phase, effective data noise reduction through curve smoothing, and applicability to data with irregular time sampling schedules (Ullah & Finch, 2013).

Dryden and Mardia (2016) primarily focused on statistical shape analysis that also discussed the foundations of landmark shape analysis, including geometrical concepts and statistical techniques that include analysis of curves, surfaces, images, and other types of object data (Dryden & Mardia, 2016). Functional data analysis considers shapes as continuous functions or curves, allowing for the analysis of shape changes over a continuum such as time or developmental stages. These studies have inspired this thesis is to investigate is the coordinates are represented in a function form via FDA approach.

Unlike traditional approaches that handle data as vectors in Euclidean space $\mathbb{R}^n$, FDA focuses on the analysis and theory of data represented as functions. In essence, each observed variable is characterised by functional values rather than discrete real values. A functional random variable is characterised by its values existing within an infinite-dimensional vector space. Functional data, in turn, represents a specific instance or realisation of such a variable. These data points are viewed as observations derived from stochastic processes operating in infinite-dimensional spaces.

The initial stage in FDA involves transforming a discrete collection of measurements, represented by observed data points, into a continuous curve, $X_1(t), X_2(t), \dots, X_n(t)$, which can either exhibit rough or smooth characteristics. Let $\Phi = \{\phi_j(\cdot) : j \in N\}$ be an infinite basis of $\mathcal{L}^2(I)$ which is the space of square integrable functions on a compact

interval of $\mathbb{R}^d, I$ with $d \in \mathbb{N}$. The elements of Φ are usually orthogonal. Every element of $\mathcal{L}^2(I)$ can be written as a linear combination of the elements of Φ. A functional random variable X valued in $\mathcal{L}^2(I)$ may be decomposed into:

$$X = \sum_{j \geq 1} c_j \phi_j(\cdot),$$

where $\{c_j\}_{j \geq 1}$ is an infinite set of coefficients (Ramsay & Silverman, 2005) . The basis expansion is used to approximate the realisation X by its projection on the span of a finite basis functions $\Phi_J = \{\phi_j(\cdot) : 1 \leq j \leq J\}$, a finite subset of Φ and $\{c_j\}_{1 \leq j \leq J}$ a subset of $\{c_j\}_{j \geq 1}$: $X \approx \sum_{j=1}^{J} c_j \phi_j(\cdot)$,

X can be summarised by a J-dimensional vector.

Functional principal component analysis (FPCA) extends the traditional multivariate PCA into the realm of functional data. Just as in the classical case, FPCA aims to achieve an optimal linear representation of a set of functional data within a finite-dimensional space. The primary objective is to diminish the dimensionality of the data through FPCA, thereby discerning the principal sources of variability (Ullah & Finch, 2013). Essentially, FPCA acts as a dimension reduction technique, reshaping the sampled curves to encapsulate the variability patterns within a lower-dimensional space. Comprehensive methodologies for FPCA are expounded upon by (Ramsay & Silverman, 2005) and (Ferraty & Vieu, 2006). For n functional observations of X in $\mathcal{L}^2(I)$, denoted as $X^{(1)}, \ldots, X^{(n)}$, J functions of $\mathcal{L}^2(I)$, $\phi_1, \ldots, \phi_J$ are sought, which are orthogonal and such that the projection of $X^{(i)}$ onto the vector space generated by these functions yield the minimum loss possible.

Principal components of FPCA that explains the variability of $\{X_i\}$ are obtained by computing the eigenfunctions corresponding to the ordered eigenvalues (from largest to

smallest) of an empirical covariance operator. Thus, performing the PCA of the X_i involves looking for the eigenvalues of the operator $\Gamma f(t) = \mathcal{L}^2(I) \mapsto \mathcal{L}^2(I)$ which is defined by:

$\Gamma f(t) = \langle C(\cdot, t), f(\cdot) \rangle, t \in I, f \in \mathcal{L}^2(I)$ and $C(s,t) = Cov[X(s), X(t)]$,

where Γ is a positive, linear, and self-adjoint operator in $\mathcal{L}^2(I)$ (Horváth & Kokoszka, 2012). It is a compact operator with a finite trace. There exists a complete orthonormal basis $\{\phi_j\}_{j \geq 1}$ and a sequence of real numbers $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ such that:

$\Gamma \phi_j = \lambda_j \phi_j$, and $\lambda_j \to 0 \; as \; j \to \infty$,

where $\{\lambda_j\}_{j \geq 1}$ is the set of eigenvalues of the covariance operator Γ associated to $\{\phi_j\}_{j \geq 1}$ the set of its eigenfunctions. The eigenfunctions corresponding to the eigenvalues are denoted as $\{\phi_j\}$. It can be shown that the eigenfunction associated with the largest eigenvalue, ϕ_1, is a solution of the following constrained optimisation problem:

$$\max_{||\phi||_2 = 1} \langle \Gamma \phi, \phi \rangle$$

where $||\phi||_2 = \int \phi^2 \, dt$ is the $\mathcal{L}^2(I)$ norm of the eigenfunction ϕ on I.

The process X can be represented using the Karhunen-Loève representation:

$$X(t) = \mu_X(t) + \sum_{j \geq 1} c_j \phi_j(t), \; t \in I,$$

where $c_j = \langle X - \mu, \phi_j \rangle$, $E(c_j) = 0$, $cov(c_j, c_l) = \lambda_j 1_{j=l}$, $\mu_X(t)$ is the mean function, $E(X(t))$ and the $\{\phi_j\}_{j \geq 1}$ are the FPCA basis. Hence, X is approximated by truncating the infinite sum at the first J terms:

$X^{(n)} \approx \mu_X(t) + \sum_{j=1}^{J} c_j \phi_j(t)$, $t \in I$ with $c_{i,j} = \langle X - \mu_X(t), \phi_j \rangle$.

In practice, since Γ is unknown, FPCA entails exploring the spectrum of the empirical covariance function:

$$\hat{C}(s,t) = \frac{1}{n}\sum_{i=1}^{n}(X_i(s) - \bar{X}_i(s))(X_i(t) - \bar{X}_i(t)),$$

where the empirical estimator of the mean of X is defined on I by:

$$\bar{X}_i(s) = \frac{1}{n}\sum_{i=1}^{n} X_i(s) .$$

MFPCA is used as a dimension reduction tool to transform sampled curves to represent the patterns of the variability of the curves, which is considered as a more natural way to represent a multivariate functional data as they share the same structure as each observation (Happ & Greven, 2018). The principal component (PC) scores obtained from both GM and FDGM are used as input to construct the linear discriminant analysis (LDA) model as it provides better classification performance (de Almeida et al., 2021). In recent years, FDA has seen applications in diverse areas such as functional neuroimaging, econometrics, and environmental science. Researchers continue to develop novel methodologies and expand the theoretical foundations of FDA to address new challenges and opportunities in analying complex functional data.

This work introduces the functional data geometric morphometrics (FDGM) approach to analyse shape variations using the functional form of the 2D and 3D landmark coordinate data. FDA is employed to analyse the image and shape data in the form of functions. Functional and shape analysis require tools to perform statistical analysis on signals, curves, or even more complex objects while being invariant to certain shape-preserving transformations (Guo et al., 2022). To ensure that the functions are well-aligned for geometric features such as peaks and valleys, curve registration (Ramsay &

Li, 1998; Srivastava et.al, 2011) or functional alignment (Ramsay, 2006) are applied to warp the temporal domain of functions (Guo et al., 2022).

Epifanio and Ventura-Campos (2011) demonstrated that FDA framework surpasses other approaches such as the landmark-based approach or even the set theory approach with principal component analysis (PCA), using a well-known database of bone outlines. FDGM treats cranial shapes of functions and curves as random variables taking values in well-defined shapes space of functions, which will help derive shape-based inferences in consideration of the geometric of the cranial shape space (Srivastava & Klassen, 2016).

The FDGM method will give a new way to observe, manipulate and use morphometrics landmark data where a data is not a value or a vector, but all available information including its dynamics. Hence, FDA is an appropriate framework to represent shapes with their intrinsically continuous or structured character whereas in multivariate GM approaches, the data are only extractions or aggregations (e.g., 3D data in GM).

This study provides a comparison of both the GM and the FDGM approach and whether the application of FDA matches or surpasses the GM method in detecting variation among biological organisms with the interest to study coordinates being represented in a function form.

2.6 Machine Learning

Machine learning (ML) encompasses a diverse array of algorithms designed to make predictions, often leveraging vast datasets (Nichols et al., 2019). In morphometric studies aimed at classification and identification tasks, the application of extensive machine learning techniques has become increasingly prevalent (Tan et al., 2018). Notably, classifiers such as naive Bayes (NB), support vector machine (SVM), random forest (RF),

and generalised linear models (GLM) are frequently employed due to their proven efficacy in numerous previous studies.

The NB classifier is grounded in Bayes' theorem, which originates from the work of Reverend Thomas Bayes in the 18th century. Bayes explored methods for computing probability distributions, particularly for binomial parameters. Although Bayesian methods have been utilised in pattern recognition for decades (Duda & Hart, 1973), they gained significant traction within the machine learning community in the 1990s. Kononenko (1990) compared the performance of inductive learning methods, specifically decision trees and the Naive Bayes (NB) classifier, for developing expert systems in four medical diagnostic problems. The study found that the NB classifier outperformed decision trees in classification accuracy, though both methods offered valuable insights into the knowledge acquired. Langley et al. (1992) conducted an average-case analysis of Bayesian classifiers, demonstrating that these classifiers perform exceptionally well on various learning tasks, particularly under the assumptions of a monotone conjunctive target concept and independent, noise-free Boolean attributes.

The NB classifier is based on Bayes' theorem and assumes that the attributes in a dataset are conditionally independent, given its class (Webb, 2011). In Rodrigues et al., (2022), NB was the best classifier for detecting landmarks in automatic wing geometric morphometrics classification of honeybee (*Apis mellifera*) subspecies. Similarly, Thomas et al. (2023) utilised NB to automate morphological phenotyping in geometric morphometrics, reducing observer bias and enhancing the capture of comprehensive representations of morphological variation.

NB is applied as one of the classifiers in this thesis due to its simplicity and computational efficiency. FDGM often involves analysing complex shapes and forms, which can be represented by a large number of features. NB can handle high-dimensional data efficiently, making it suitable for quick, initial analysis or as a baseline model.

SVM addresses a multi-class problem as a single "all-together" optimisation. This classifier can be used to find a hyperplane in a 2-dimensional space that will separate the scores to their potential species. Bellin et al. (2021a) successfully combined GM with different machine learning algorithms, including SVM with radial basis function (RBF) kernel. This study demonstrated the effectiveness of SVM in correctly classifying two *Anopheles* sibling species of the *Maculipennis* complex based on shape data (Bellin et al., 2021). Motivated by such successes, this study aims to leverage supervised learning, particularly SVM, for the classification of shrew species and dietary of kangaroos based on their morphological features. As morphometric data can be complex and prone to overfitting, SVM's regularisation techniques can be useful to avoid overfitting especially when the number of features is large relative to the number of samples.

RF, a classification algorithm developed by Breiman (2001) based on bootstrap aggregating or bagging that combines the predictions of multiple decision trees to make a final prediction. Breiman (1996) introduces the concept of bagging (which is a fundamental idea used in RF. It describes how combining multiple models can enhance predictive performance. Arai et al. (2021) applied RF in the context of morphological identification in skulls, specifically between spotted seals and harbor seals, using GM. The study achieved an identification accuracy rate of 100% using RF by narrowing down to a subset of eight key landmarks out of a total of 75 landmarks (Arai et al., 2021). The ensemble nature of RF allows it to capture both linear and non-linear relationships in the data, making it robust and accurate for shape classification tasks.

The success of RF in morphological identification (Bellin et al., 2021a; Berio et al., 2022; Khang et al., 2021) has encouraged this study to compare the effectiveness of this classifier in the classification of the shrew species and dietary of kangaroos based on the FDGM framework. GLMs, as extensions of linear models, offer flexibility in accommodating nonlinearity and non-constant variance within data distributions.

Consequently, GLMs are well-suited for analysing species-habitat relationships, which often exhibit deviations from normal distributions (Chiaverini et al., 2023).

ANN models use neural networks, which are based on the understanding of the biological nervous system. These models 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). McCulloch and Pitts (1943) introduced the concept of artificial neurons and their ability to perform logical operations, laying the groundwork for later developments in neural networks. Rosenblatt (1958) introduced the Perceptron, an early type of neural network used for binary classification. This work was crucial in demonstrating that neural networks could learn and make decisions based on input data. Rumelhart et al. (1986) presented the backpropagation algorithm for networks of neuron-like units. The procedure repeatedly adjusts the weights of the connections in the network so as to minimise a measure of the difference between the actual output vector of the net and the desired output vector. As a result of the weight adjustments, internal 'hidden' units which are not part of the input or output come to represent important features of the task domain, and the regularities in the task are captured by the interactions of these units, thus significantly advancing the field of deep learning. Rojas (1996) provides a comprehensive overview of neural networks, including the development and application of multi-layer perceptrons. This book is a key reference for understanding the evolution of neural network models.

ANN are inspired from the human brain that works as a paradigm to perform computations in an effective and efficient manner (Mas & Flores, 2008). In a study by Salifu et al. (2022), RF, SVM and ANN were also evaluated for their predictive performance in discriminating fruit fly species. The study concluded that SVM and ANN models outperformed RF in accurately classifying fruit fly species. ANN can be useful in capturing complex and non-linear relationships between morphological features that

might be missed by simpler linear models through their multi-layer structure and non-linear activation functions. This makes ANN a powerful tool for analysing high dimensional morphometric data, making them well-suited for a wide range of applications in FDGM. Inspired by this study, this research explores the predictive performance of these models across different biological organisms.

CHAPTER 3: RFE-BASED FEATURE SELECTION TO IMPROVE CLASSIFICATION ACCURACY FOR TRADITIONAL MORPHOMETRIC ANALYSIS

3.1 Introduction

In the field of machine learning, the selection of relevant features is crucial for enhancing model performance and accuracy. This study incorporates the application of the recursive feature elimination (RFE) method to select pertinent features from the craniodental linear measurements of male and female *Rattus rattus*, a rodent species native to the Indian Peninsula and a common pest in Malaysia. By refining the feature set, the study aims to improve the learning performance and classification accuracy of predictive models, thereby contributing to more effective data-driven solutions in rodent pest management.

Feature extraction and feature engineering are foundational processes in machine learning, involving the creation of new features from existing ones based on domain-specific knowledge. This process increases the number of features available for analysis, which is essential for capturing more nuanced patterns within the data. However, before these features can be effectively utilised, a selection process must be undertaken to identify the most informative subset. Initially, feature extraction generates a broad array of potentially useful features. Subsequently, feature selection narrows this down to the most impactful ones, thereby enhancing the model's performance.

Dimensionality reduction is another critical concept in this context. While it shares the goal of reducing the number of features with feature selection, the methods differ significantly. Feature selection involves retaining a subset of the original features and discarding the rest. In contrast, dimensionality reduction projects the original features onto a lower-dimensional space, creating a new set of features. Practically, either

approach can be used, but when both are applied, feature selection should precede dimensionality reduction to streamline the dataset effectively.

Feature selection is driven by several key considerations that collectively enhance the efficiency and efficacy of machine learning models. Firstly, features that have no relationship with the target variable can introduce noise, leading to overfitting. Removing these irrelevant features helps maintain model robustness. Additionally, redundant features, even if important, can be discarded if another feature encapsulates their information. This mitigates issues such as multicollinearity, particularly in linear models.

High-dimensional datasets can suffer from the curse of dimensionality, where each data point becomes sparse, making it difficult for the model to learn meaningful patterns. Feature selection reduces dimensionality, thereby enhancing the model's learning capability. Moreover, models with too many features often lose interpretability. Simplifying the feature set improves the model's interpretability, which is particularly important in regulated domains where interpretability may be a legal requirement.

RFE is a powerful technique that aligns closely with backward selection but differs in its execution. While backward selection relies on a model performance metric from a hold-out set, RFE eliminates features based on their importance as determined by the model itself. This importance can be derived from feature weights in linear models, impurity decrease in tree-based models, or permutation importance applicable across various model types. By iteratively removing the least important features, RFE refines the feature set to enhance model performance.

The black rat, *Rattus rattus* Linaeus, 1958 is a widespread rodent pest with significant ecological and economic impacts. In Malaysia, research on *R. rattus*, especially regarding feature selection techniques for craniodental measurements, remains limited. This study addresses this gap by employing NB, RF, and ANN as predictive models to classify age groups of both male and female *R. rattus*. The performance of these models, utilising

RFE-selected variables, was compared and analysed.

The application of RFE in this study demonstrates its utility in refining feature sets to improve model performance. By selecting the most relevant craniodental measurements, the predictive models achieved higher classification accuracy, underscoring the importance of feature selection in machine learning. This approach not only enhances the effectiveness of predictive models but also contributes to better understanding and management of *Rattus rattus* populations.

3.2 Methodology

3.2.1 Data Description of the *Rattus rattus* Data

A total of 130 individuals of *R. rattus* were caught and examined for skull morphometrics study. The male and female *R. rattus* cranial and mandible measurements (67 males and 63 females) were used in this study i.e., 20 morphometric variables (see Mohamad Ikbal et al. (2019)). Figure 3.1 and Table 3.1 show the parts of measurements taken based on Musser & Newcomb (1983) and Musser, et al. (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 (15 males and 24 females); C3: The longitudinal link between the first and second lobes of the upper M3 is larger and generally wider all the linear measurements in the dataset and the results are tabulated (16 males and 14 females); 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 (36 males and 25 females).The original dataset of the linear measurements and age classes were then split for each sex using the 70/30 test/train split (70% of the whole dataset used for training, and 30% for testing) based on random sampling across combination of age

for both male and female rats before fitting it to the RFE model to prevent overestimation of accuracy in the empirical analysis.

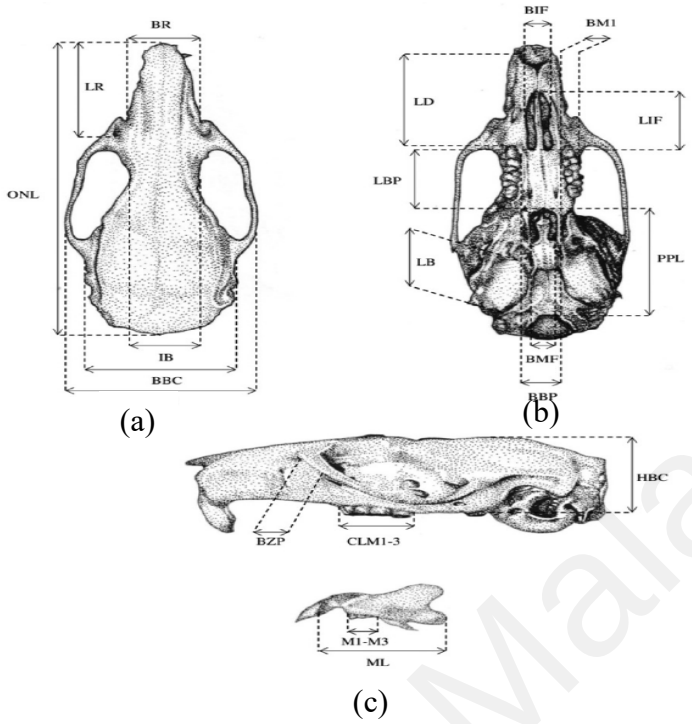


Figure 3.1: Craniodental measurements of *R. rattus* based on the (a) dorsal, (b) ventral, and (c) lateral views (Photo sourced from Muhammad Ikbal et al., 2019)

Table 3.1: Model performance evaluation based on age groups for male *R. rattus*

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)		

Table 3.2 The localities and samples sizes from which *R. rattus* populations were collected in Peninsular Malaysia.

Locality	**Sample size**	**Habitat**
Kuala Perlis, Perlis	9	Seaside
Kota Bharu, Kelantan	8	Housing area
Alor Setar, Kedah	12	Housing area
Georgetown, Penang Island	6	Seaside
Seberang Jaya, Penang mainland	7	Housing area
Kuala Terengganu, Terengganu	10	Housing area
Ipoh, Perak	8	Fresh market
Kuantan, Pahang	12	Housing area
Chow Kit, Kuala Lumpur	15	Fresh market
Shah Alam, Selangor	9	Housing area
Seremban, Negeri Sembilan	12	Fresh market
Masjid Tanah, Melaka	10	Housing area
Stulang Laut, Johor	12	Seaside

3.2.2 Recursive Feature Elimination

In the field of machine learning, selecting the right features is crucial for building efficient and accurate models. While decision trees are popular for feature selection due to their simplicity and interpretability, the RFE method offers several advantages that make it a compelling alternative. RFE is an effective feature selection method that initially uses the entire set of features to build the model. This feature selection technique can be applied to any model that can rank features by importance, such as support vector machines (SVMs), linear models, and random forests. This flexibility allows for a broader application across different types of machine learning algorithms. Decision trees are a tree-structured model used for both classification and regression tasks. For feature selection, they rank features based on their ability to split the data into homogeneous subsets, often using metrics such as Gini impurity or information gain. RFE's ability to work with a variety of models (e.g., SVMs, linear models) provides greater flexibility compared to the decision tree method, which is inherently tied to the tree structure. RFE's iterative approach ensures that the feature selection process is thorough and optimised for the final model's performance. While decision trees offer simplicity and interpretability in feature selection, RFE provides a more flexible and robust approach, particularly suited for improving model performance and generalisation.

Since RF deals well with high dimensional data problems (Darst et al., 2018), this algorithm was applied 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 (Misra & Singh, 2020). Based on the accuracy of different attribute subset sizes obtained, the top performing features from the RFE model were then chosen 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 & Singh, 2020).

3.2.3 Classification Models

Morphometric studies for classification and identification tasks are enhanced by extensive machine learning methods. The naive Bayes (NB), random forest (RF), and artificial neural network (ANN) classification models are frequently applied because they have been successfully used in many previous studies.

The NB classification model is a classifier which provides a mechanism that utilises predictors of the training data to estimate the posterior probability, $P(y_k|\,\mathbf{x})$ Sammut & Webb, 2010). NB classifiers were trained 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:

$$P(y_k|\,\mathbf{x}) = \frac{P(y_k)P(\mathbf{x}|\,y_k)}{P(\mathbf{x})}$$

where $\mathbf{x}$ represents the scaled linear measurements and y_k represents the age classes of the rats' training data. $P(y_k)$ is the prior probability of class y_k. Given the age classes are C2, C3 and C4 for *R. rattus*, the classification problem is formulated as a multiclass classification problem because there are more than two classes. Under the NB assumption, the features are conditionally independent given the class. Therefore, the likelihood $P(\mathbf{x}\,|\,y_k)$ can be expressed as the product of the individual conditional probabilities:

$$P(\mathbf{x}|\,y_k) = \prod_{i=1}^{n}(x_i|\,y_k).$$

The NB classifier assigns the individual to the age class y_k with the highest posterior probability:

$$y = \underset{y_k}{\operatorname{argmax}} P(y_k) \prod_{i=1}^{n} (x_i | y_k).$$

Random Forest (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). The RF model with all the predictor variables of the training data was fitted 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. RF offers a different approach to machine learning compared to NB and ANN. While NB is a probabilistic classifier based on Bayes' theorem and ANN is a biologically inspired model that learns from data, RF is an ensemble learning method based on decision trees. Including RF allows for a more comprehensive comparison across different machine learning paradigms.

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, all features of the training data were applied into the neural networks and select the age classes of rats as targets. The architecture of the ANN used is as follows:

(a) Input Layer: The input layer receives the initial data that need to be processed which is represented as neurons. Each neuron corresponds to one feature and this layer passes the information on to the next layer in the ANN. In this study, the linear measurements of the male and female rats were used as the input layer.

(b) Hidden Layers: These layers are intermediate layers between the input and output layers. Actual computation of input data is performed in each neuron of a hidden layer based on the input received from the neurons in the previous layer using a weighted sum followed by an activation function to produce an output. Activation function is a critical component of a neural network that introduces non-linearity into the model so that the network is able to learn complex patterns and relationships in the data. In this study, using method "nnet" for neural networks, the default activation function used is the logistic sigmoid function.

(c) Output layers: The output layer provides the final results of the neural network computation. The number of neurons in the output layer depends on the nature of the task. In this study, the number of neurons in the output layer corresponds to the number of age classes of the male and female rats (C2, C3 and C4).

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.

3.2.4 Performance Evaluation Metrics for Classification Models

The multiclass confusion matrices of the classification models were observed and their performances between the models were compared 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 & Singh, 2020), i.e., 29.23% are C2, 23.08% are C3 and 47.69% are C4, Kappa, precision, recall and F1 score measures were observed to evaluate the performance of the machine learning algorithms.

These measures are calculated for each age class as follows:

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

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

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

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

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.

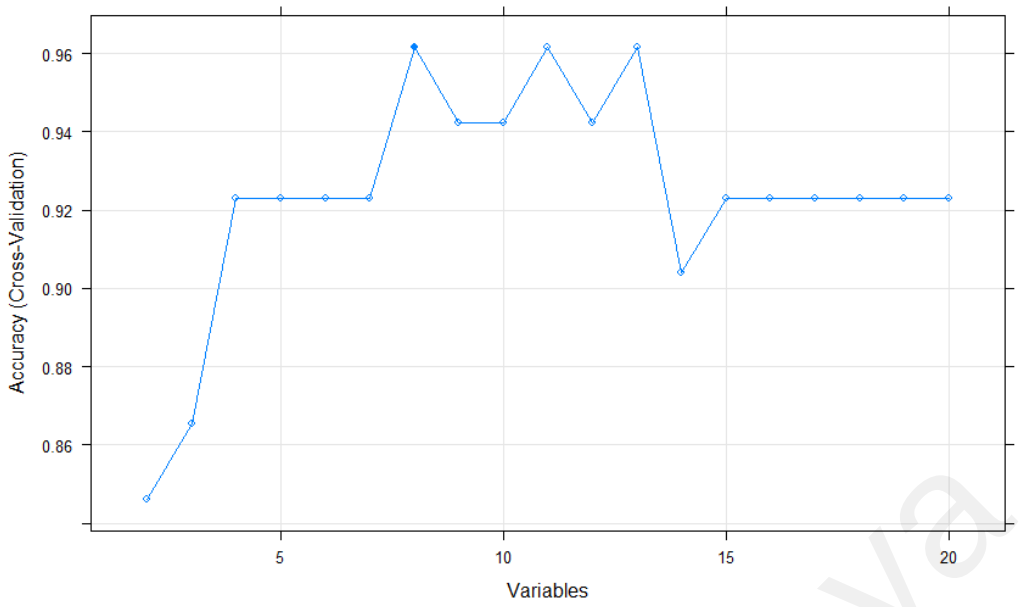
All statistical analyses were performed using R. The caret package (Kuhn, 2008) was used in R version 4.2.1 (R Core Team, 2023) to apply the RFE algorithm and to streamline the model training process for classification tasks. In addition, the factoextra package (Kassambara & Mundt, 2020) and ggfortify package (Tang et al., 2016) were also applied in R to visualise the PCA output. The santaR package (Wolfer et al., 2022) is used to scale the linear measurements of both male and female rats at unit variance. The MLeval (Christopher & John, 2022) package is applied to construct the ROC curves for the classification models with all features and RFE-selected features.

The default hyperparameters are used for all three machine learning models trained using the caret package. The 'trainControl' function is utilised to define the resampling method and parameters, which specifies 10-fold cross validation. For the NB model, the default settings include a kernel density estimate for continuous variables (by default, kernel density estimation is not used), a smoothing parameter for the conditional probability tables (by default, no smoothing is applied), an adjustment factor adjust for bandwidth in kernel density estimation (by default, no adjustment is used) and a cut-off for classification of 0.5, by default. The hyperparameters for the RF model includes the number of variables randomly sampled as candidates at each split (by default, it is the square root of the number of variables), the number of trees is 100, minimum size of terminal nodes (by default is 1 for classification). For the ANN model, the hyperparameters are the number of units in the hidden layer(s) (by default, it is 1), a decay term for weight decay (by default is 0, meaning no decay), maximum number of iterations, which is 100, the maximum number of weights (default is 0, meaning unlimited), and no entropy error is used by default.

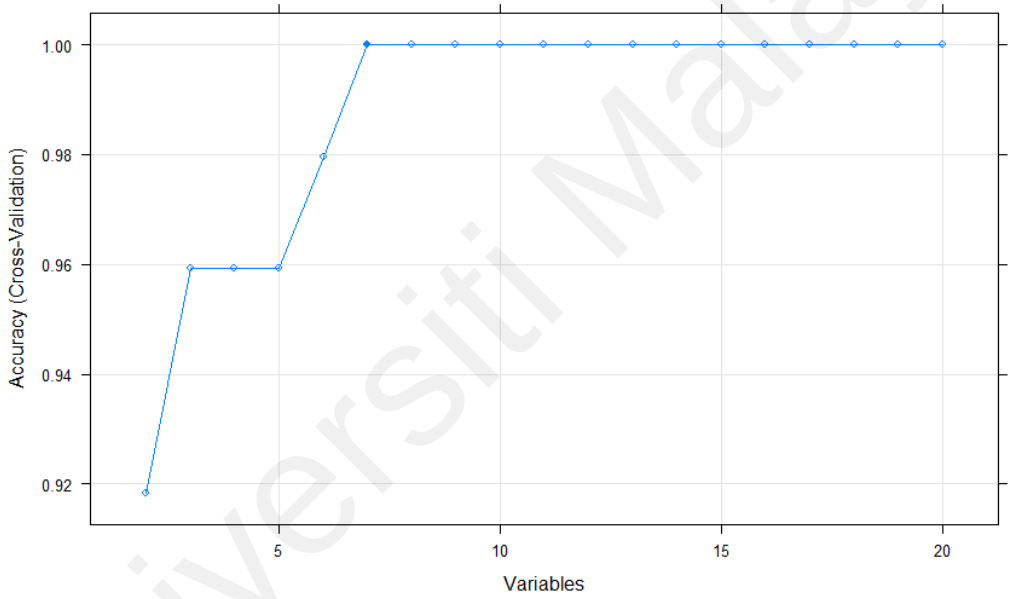
3.3 Results and Discussion

After performing the train-test split for the skull measurements data of both male and female rats, the automatic RFE was applied, 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 shown in Figure 3.2, HBC, IB, LD, BZP, BR, ZB, and LIF were identified by the RF-RFE as the features that indicate significant differences among age classes for *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 (Alamoudi et al., 2021).

Male rats of this genus tend to have longer rostrum with shorter and wider zygomatic arch (Alamoudi et al., 2021). As for *R. rattus* females, the top performing features to distinguish age classes are ZB, LD, BMF, BBC, IB and BR. The top performing features were obtained using the "rfe" function in the library "caret". 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 et al., 2021). All features selected by RFE for the male and female rats appear to coincide with most of the craniodental measurements used in Balakirev et al. (2011); Breno et al. (2011); Esselstyn et al. (2015); Libois et al. (1996); Motokawa et al. (2004); Timm et al. (2016). Among all features selected by RFE, the zygomatic breadth (ZB), interorbital breadth (IB), breadth of rostrum (BR) and length of diastema (LD) were observed to be significant in both male and female rats, which is 44.4% of the "total chosen features" of both male and female rats. These selected features are later used in PCA, LDA and the predictive classification models for each sex and their performance measures are evaluated and tabulated.



(a)



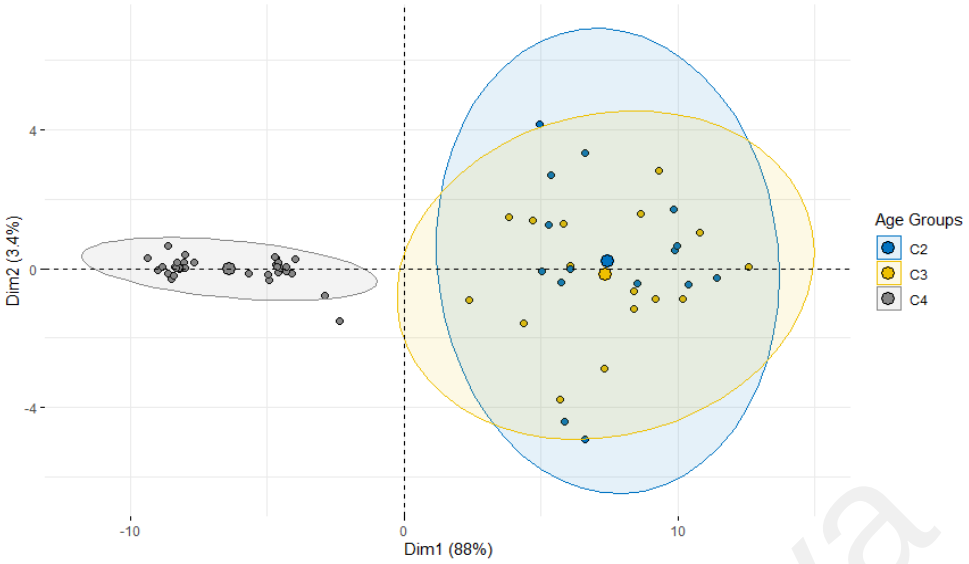
(b)

Figure 3.2: Performance profile plots across different subset sizes given by RFE approach for scaled (a) male and (b) female craniodental measurement dataset

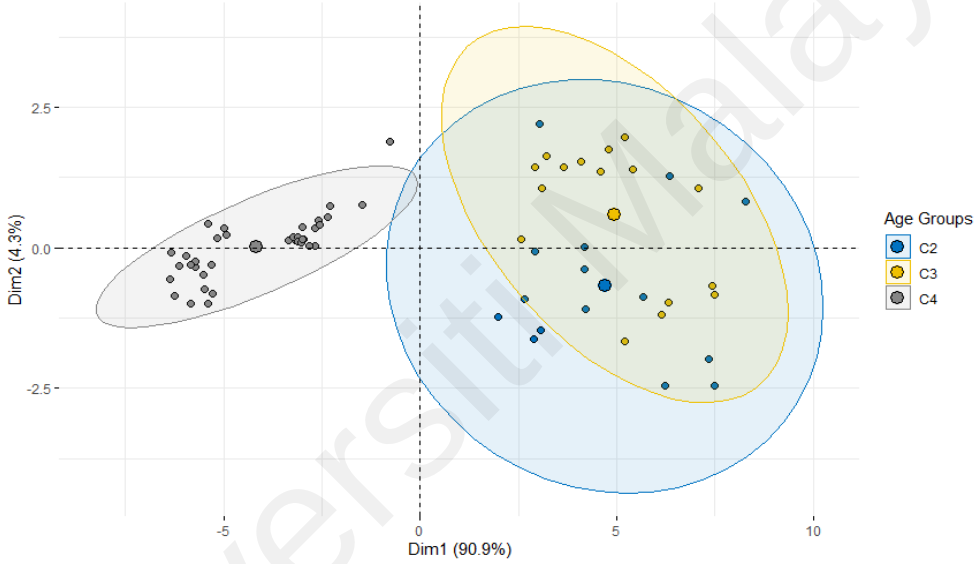
3.3.1 Principal Component Analysis

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 (Figure 3.3(b)(i). 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 reveals more distinct clusters among the age groups in female rats (Figure 3.3(b)(ii)). 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 Mohamad Ikbal et al. (2019) with 14 of the craniodental measurements for both sexes.

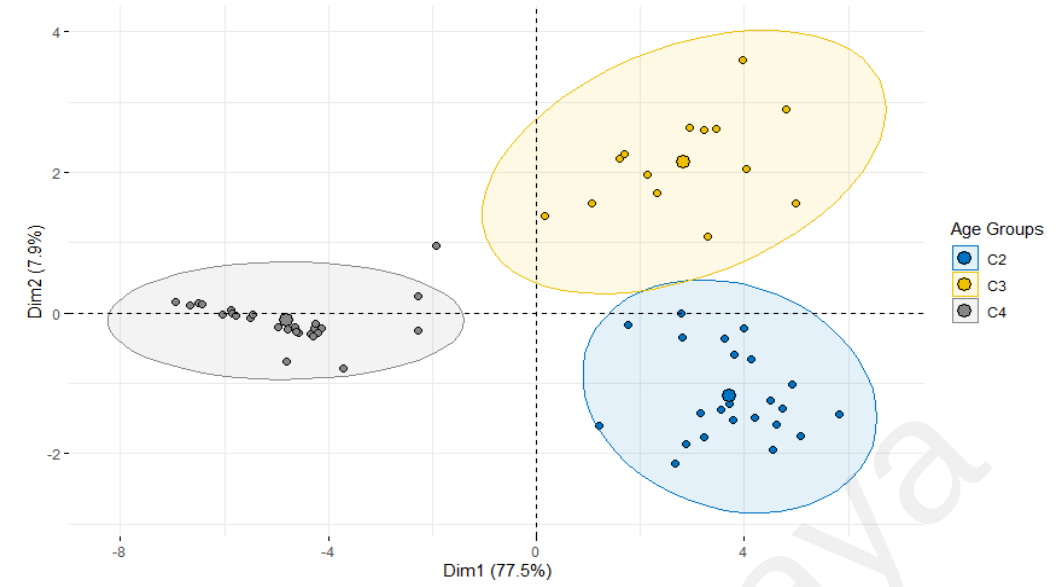


(i)

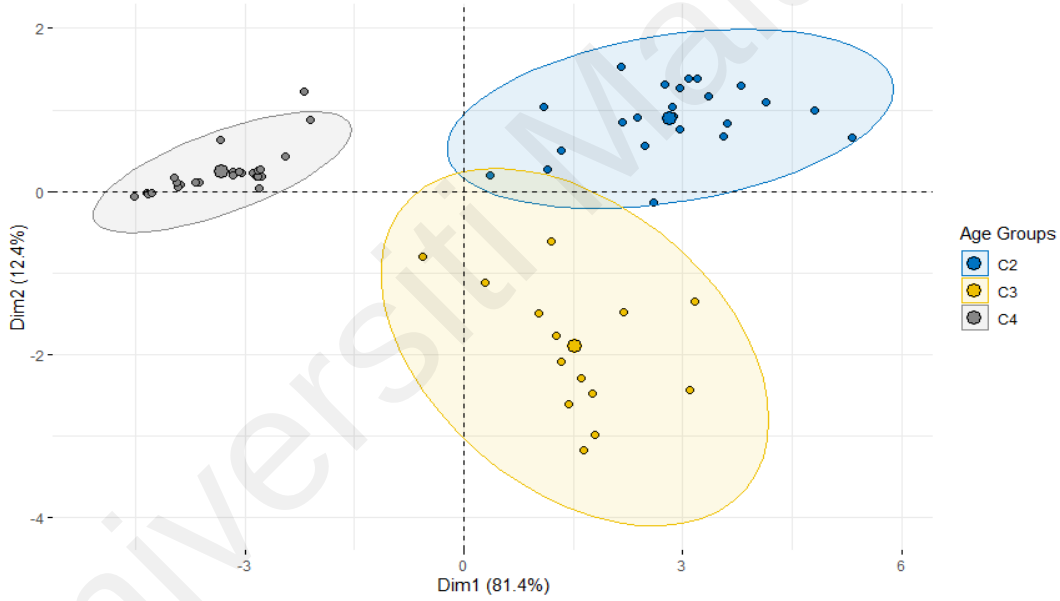


(ii)

Figure 3.3: PCA plots for *R. rattus* male craniodental measurement ((i) all features (ii) significant features. The ellipses help visualise the spread and central tendency of each group. Each ellipse encompasses 95% of the individuals within that group, indicating where most of the data points for each group are concentrated



(i)



(ii)

Figure 3.4: PCA plots for *R. rattus* female craniodental measurement ((i) all features (ii) significant features. The ellipses help visualise the spread and central tendency of each group. Each ellipse encompasses a 95% of the individuals within that group, indicating where most of the data points for each group are concentrated.

3.3.2 Predictive Classification Models Performance

The RFE selected features are used in the predictive classification models for each sex and their performance measures are evaluated and tabulated (Table 3.2 and Table 3.3).

Based on Table 3.2 and Table 3.3, both training and test sets give excellent results in terms of accuracy for all three models (evaluated once), with all features included, where ANN is the best model for both male and female craniodental measurement datasets.

A comparable result is observed in the NB model for *R. rattus* males after fitting the top performing features into the model, which indicates that RFE can be considered as an alternative feature selection method. 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. Class imbalance can lead to biased models that perform poorly on the minority class. 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 3.4).

Based on the age groups of the male *R. rattus,* it was observed that all the three models yield high scores for precision (Table 3.3). 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 reveal 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 3.4) 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.

ANN was chosen 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.

Table 3.3 ROC-AUC results for *R. rattus* male and female craniodental measurement ((i) all features (ii) top performing features)

Classifiers	**Male rats**		**Female rats**	
	All features	RFE features	All features	RFE features
NB	0.96	1.00	1.00	0.98
RF	0.97	0.98	0.98	1.00
ANN	0.98	1.00	1.00	1.00

Table 3.4: Model performance evaluation based on age groups for male *R. rattus*

Classifiers	**Training data accuracy**		**Test data accuracy**		**Kappa**	
	All features	RFE features	All features	RFE features	All features	RFE 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 3.5: Model performance evaluation based on age groups for female *R. rattus*

Classifiers	Training Data Accuracy		Test Data Accuracy		Kappa	
	All features	RFE features	All features	RFE features	All features	RFE 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

Table 3.6: Precision, recall and F1 scores of classification models using RFE-selected features for male *R. rattus* based on age groups

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

Table 3.7: Precision, recall and F1 scores of classification models using RFE-selected features for female *R. rattus* based on age groups

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

3.4 Conclusion

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. It is of interest of this thesis to examine how well RFE works when incorporated with PCA in morphometric studies. Based on the analysis for *R. rattus* males and females, a comparable result was noticeable on the performance metrics of the three predictive classification models and in PCA when the RFE-selected features are used. ANN outperforms the other models for both sexes. It was 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 Mohamad 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. Although RFE is a valuable technique for identifying relevant features, it may result in the selection of features that are correlated with each other. In this study, RFE was instrumental in identifying the most informative features for constructing predictive models, thereby enhancing model interpretability and potentially mitigating overfitting.

CHAPTER 4: FDGM IN 2D GEOMETRIC MORPHOMETRICS

4.1 Introduction

The study of craniodental morphology in shrews stands out as an invaluable avenue for gaining insights into their evolutionary trajectory, taxonomic classification, and ecological adaptations. Shrews, belonging to the order *Eulipotyphla* are characterised by their small size, insectivorous diet, and rapid metabolism. Despite their small stature, shrews exhibit remarkable diversity in craniodental morphology, reflecting adaptations to different ecological niches and evolutionary pressures. This is evident in the study conducted by (Vasil'ev, & Kourova (2015) which revealed geographical variability of the shape of mandible in three shrew species of genus *Sorex* using GM. Notably, discriminant analysis of Procrustes coordinates derived from the GM method enabled high percentage of correct assignment of individual shrews to distinct local taxocenes, further validating the efficiency of this methodology in taxonomic studies. Moreover, findings by (Vilchis-Conde et al. (2023) reinforce the significance of GM in supporting the taxonomic classification of semifossorial shrews. The research also revealed that the shapes of the skull, particularly the dentary has associated with the diet specialisation, highlighting the profound impact of morphological variations on functional aspects such as bite force among shrews. This thesis focuses on the craniodental variation among three shrew species: *Crocidura malayana* Robinson & Kloss, 1911, *Crocidura monticola* Peters, 1870 and *Suncus murinus* (Linnaeus, 1766).

Each species occupies distinct ecological niches: *C. malayana*, a medium-sized shrew, thrives in Thailand, Malaysia, and several offshore islands (Hutterer, 2005). This terrestrial species has been documented in both hill and lowland forests (Francis, 2008; Jamaluddin et al., 2022).

Meanwhile, *C. monticola*, the smallest shrew in the genus *Crocidura* is restricted to forest areas in Malaysia and Indonesia (Omar et al., 2013). On the other hand, *S. murinus*, the largest shrew species, is predominantly found in urban areas and the outskirts of forests, with a wide distribution spanning human settlements in the Indian subcontinent and Southeast Asia (Ruedi et al., 1996).

In this thesis, 90 adult shrew specimens were collected, with 30 individuals from each species. The habitats of *C. malayana* span diverse locations, including Lata Belatan, Terengganu; Ulu Gombak; Aur Island, Johor; Pangkor Island, Perak; Bukit Rengit, Pahang; Cheras Road, Kuala Lumpur; Port Dickson, Negeri Sembilan; and Dusun Tua, Selangor. Conversely, *C. monticola* exhibits a broader habitat range, inhabiting environments such as Ulu Gombak; Wang Kelian, dominated by secondary lowland forest, and Maxwell Hill, an upper dipterocarp forest, among others. Suncus murinus, on the other hand, is observed in locations like Wang Kelian, Perlis; Alor Setar, Kedah; Air Hitam, Pulau Pinang; Lumut, Perak; Ulu Gombak, Selangor; and Bukit Katil, Melaka. These varied habitats likely contribute to the divergence in craniodental morphology between species. Notably, *C. malayana* and *C. monticola* coexist in sympatry in Ulu Gombak, sharing the same habitat or niche. This study aims to elucidate the relationships between these species, offering valuable insights into the evolutionary processes shaping their craniodental morphology.

FDA is a statistical methodology used to analyse data that are represented in the form of functions, consisting of entire curves or other continuous functions, rather than discrete observations. Functional data analysis is particularly useful when dealing with data that vary continuously over a domain, such as time, space, or wavelength.

In the context of this study, the basic idea behind FDA is to express discrete observations, i.e., landmark coordinates, in the form of a function (to create functional data) that represents the entire measured function as a single observation, and later generate models to predict information based on a collection of functional data by applying statistical concepts from multivariate data analysis (Ullah & Finch, 2013).

In this work, the FDGM method is employed to analyse the image and shape data in the form of functions. The landmarks obtained from the craniodental shapes of three species of shrews are represented in the form of functional data. This data is used to perform multivariate functional principal component analysis (MFPCA) to observe variation among the three shrew species and compared with the classical PCA. The principal component scores obtained from MFPCA (MFPC scores) captures the major sources of shape variation among the shrew species. These MFPC scores are then reconstructed based on a truncated multivariate Karhunen-Loeve representation to produce predicted functions, thus allowing for a compact representation of the functional data. The results of this study revealed that FDA can be used to identify subtle differences in shape, and it can be used to relate these differences to underlying factors, such as ecology or behavioral factors.

In this study, the landmark coordinates used in the GM method will be represented as functions. Each sample element is considered as a function under the FDA framework which often defines time, spatial location, or wavelength as the physical continuum. Functional data geometric morphometrics (FDGM) is proposed in this study, requiring steps to perform statistical analysis on signals, curves, or even more complex objects while being invariant to certain shape-preserving transformations. To ensure that the functions are well-aligned for geometric features such as peaks and valleys, curve registration or functional alignment are applied to warp the temporal domain of functions.

The FDA framework surpasses its counterparts, including both the landmark-based approach and the set theory approach with principal component analysis (PCA), when applied to a well-known database of bone outlines. The set theory approach is adopted from a methodology outlined in Horgan (2000), treating shapes as sets. Each position within the image corresponds to a binary variable, indicating whether it belongs to the shape or not. Consequently, the study performed PCA specifically tailored for binary data.

The landmarks obtained from the craniodental shapes of three species of shrews are represented in the form of functional data. This data is used to perform multivariate functional principal component analysis (MFPCA) to observe variation among the three shrew species and compared with the classical PCA. The principal component scores obtained from MFPCA (MFPC scores), which capture the major sources of shape variation among the shrew species. The functional data of landmarks sampled from studied curves were then concisely represented by a continuous curve based on Karhunen-Loeve theorem. The results of this study revealed that FDGM can be used to identify differences in shape by classification methods. These differences can be used to relate to underlying factors such as ecology or behavioral factors.

This work aims to introduce geometric morphometrics in a functional data framework to reveal the existence of significant differences in craniodental shapes of three species of shrews. These differences are related to the different ecological niches that these three species occupy. The results of this study will provide valuable insights into the morphological variation among shrews. This information could be used to improve our understanding of the evolution of shrews and to develop new methods for identifying and classifying shrews.

4.2 Data Description

4.2.1 Shrew Skull Image Acquisition

The skulls of *C. malayana*, *C. monticola*, and *S. murinus* were examined from various angles, including dorsal, jaws, ventral, and lateral views (Figure 4.1). However, the ventral view was excluded from this study because it is identical to the dorsal view (Abu et al., 2018).

Ninety specimens of the three shrew species (30 for each species) were obtained from the Museum of Zoology at Universiti Malaya (UM) in Kuala Lumpur, Malaysia. The skulls from each specimen were individually placed in small bottles for GM analysis. Digital images of the skulls were captured following the method outlined by Abu et al. (2016) using a Nikon D90 camera with 15x magnification. The images were saved in Tagged Image File Format (TIFF) at a resolution of 4288 × 2848 pixels. Adobe Photoshop CS6 was used to enhance the image quality.

4.2.2 Landmark Data Acquisition

After acquiring the images, TPSUtil32 (Rohlf, 1990) is used to obtain the TPS files for all three views which will be used in TPSDig2 (Rohlf, 1990) for landmarking. Each craniodental view has different numbers of landmarks and semi-landmarks, i.e., dorsal (25 landmarks), jaw (50 landmarks) and lateral (47 landmarks). The statistical analysis of three views was performed in R version 4.2.1. To use the GM data, the raw coordinates obtained from the landmarks of all three craniodental views were processed using GPA for optimal registration using translation, rotation, and scaling using the *gpagen* function in the *geomorph* package (Adams et al., 2013). According to McCane (2013), outline methods produce useful and valid results when suitably constrained by landmarks. This leads to the main idea of this work to incorporate FDA approach to observe the separation among the three shrew species.

After the images are acquired, TPSUtil32 is used to obtain the TPS files for all three views. These files will be used in TPSDig2 for landmarking. A repeated measurement approach was employed. This approach involved having the same observer measure the outlines three times to assess the consistency and reproducibility of the measurements. By comparing these repeated measurements, any variation or error introduced by the observer during the process could be quantified and evaluated. The average of these repeated measurements was used for further analysis.

For the dorsal view, 25 landmarks were placed including 16 Type I landmarks (LM1, LM4-LM11, LM13-LM15, LM22-LM25) and 9 Type III landmarks (SLM2-SLM4, SLM12, SLM16-SLM21). Similarly, in the jaw view, 50 landmarks were positioned, comprising 32 Type I landmarks (LM1, LM3-LM22, LM24-LM26, LM32-LM35, LM41-LM43, LM48, LM50) and 18 Type III landmarks (SLM2, SLM23, SLM27-SLM31, SLM36-SLM40, SLM44-SLM47, SLM49).

Lastly, the lateral view consisted of 40 landmarks being Type I (LM1, LM4-LM11, LM15-LM18, LM20, LM22-LM47) and 7 landmarks being Type III (SLM2, SLM3, SLM12-SLM14, SLM19, SLM21).

As suggested by MacLeod (2013), the application of any specific treatment to semi landmarks, such as the sliding landmark analysis for geometric morphometric analysis has been refrained from this study. This is to prevent any alteration of the original geometric relationships which would complicate the interpretation of the results.

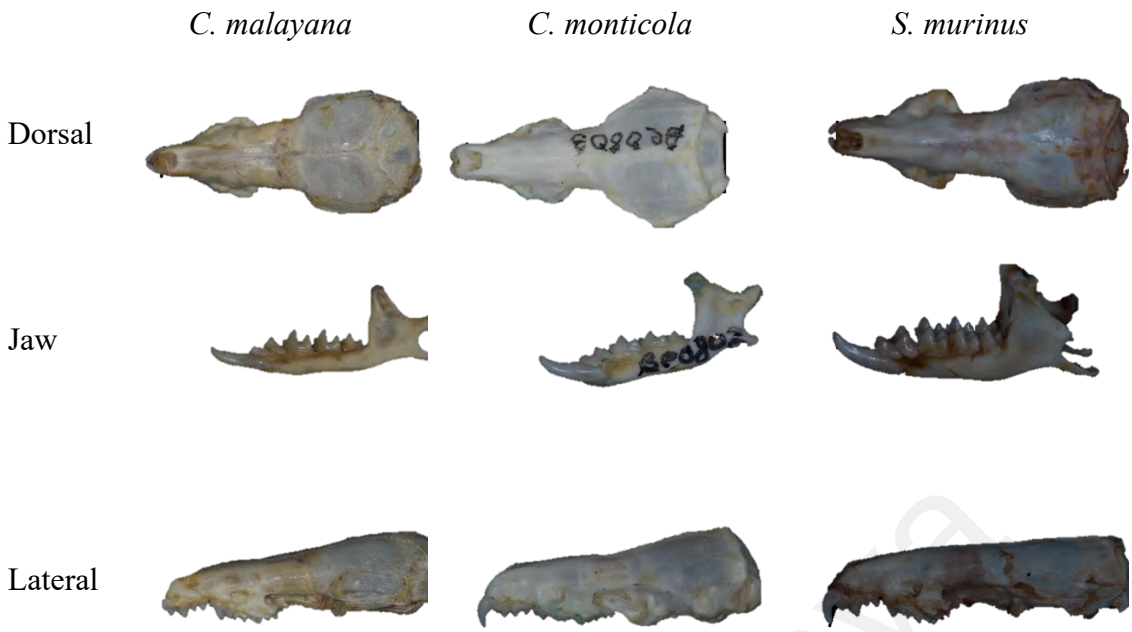


Figure 4.1: Digital skull images of dorsal, jaw and ventral views of *C. malayana, C. monticola and S. murinus*.

4.3 Functional Data Geometric Morphometrics in 2D Landmark Data

4.3.1 Functional Landmark Data

This thesis introduces functional data framework of geometric morphometrics known as functional data geometric morphometric (FDGM). In this framework, FDA is integrated with GM to capture and analyse shape variations across specimens. This integration allows for a more comprehensive analysis of shape variations by considering landmark coordinates as functional data. FDA is a method used to analyse raw data that varies dynamically over time, space, or more complex dimensions. In this study, standardised coordinates from GPA were employed to evaluate the outlines of the shapes in three craniodental views. As the methodology is similar to that in Chapter 5, the FDGM method is shown using 3D landmark representation. Each observation is vector-valued, as three spatial coordinates which are the x, y and $z-$ coordinates are involved.

Let

$$\left\{ \left(x_k(t_1), \mathrm{y_k}(t_1), \mathrm{z_k}(t_1) \right)^{\mathrm{T}}, \ldots, \left(x_k(t_\tau), \mathrm{y_k}(t_\tau), \mathrm{z_k}(t_\tau) \right)^{\mathrm{T}} \right\};$$

where $k = 1, \dots, n$ be the standardised landmark coordinates for n specimens and $\tau = 1, \dots, p$ be the number of landmarks on a $d-$ dimensional domain (example of 2D representation can be referred to Figure 4.2 (a); Figure 4.3(a); Figure 4.4(a)). To implement functional data in an object-oriented way, the raw data is converted into functions.

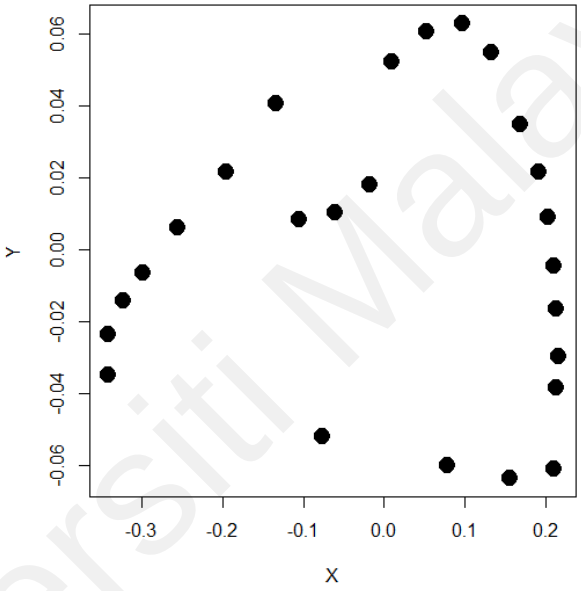
To mitigate non-shape variations such as translation, rotation, and scaling, Procrustes superimposition is employed on landmark coordinates in each view. This ensures alignment of landmarks while preserving shape differences across specimens. (Figure 4.2 (b); Figure 4.3(b); Figure 4.4(b)). This work is inspired by the study conducted by Happ-Kurz (2020) and is based on the crania of the shrews.

Let $\{x_k(t_1), \dots, , x_k(t_\tau)\}$, $\{y_k(t_1), \dots, , y_k(t_\tau)\}$ and $\{z_k(t_1), \dots, , z_k(t_\tau)\}$ where $k = 1, \dots, n$ be the separated standardised landmarks for n specimens for x, y and $z-$ coordinates respectively. The data is organised in two fields to facilitate FDA in an object-oriented manner. For example, the $x-$coordinates are used as the observation points (boundaries) $\{t_{k1}, \dots, t_{kp} : k = 1, \dots, n\}$ and the values of landmarks represent the set of observed values $\{x_{k1}, \dots, x_{kp} : k = 1, \dots, n\}$. This creates a data block of a univariate functional data object, representing the $x-$coordinates as a collection of vectors that define the marginals of the observation grid (Happ-Kurz, 2020). The same process is applied to the y and $z-$ coordinates.

These discrete curve observations were converted into continuous functions, $X_k(t)_{k=1,\dots n}$, $Y_k(t)_{k=1,\dots n}$ and $Z_k(t)_{k=1,\dots n}$using the *funData* package (Happ & Greven, 2018) in R. This approach represents the landmark points as univariate functional data with n observations as a list for x and y- coordinates respectively.

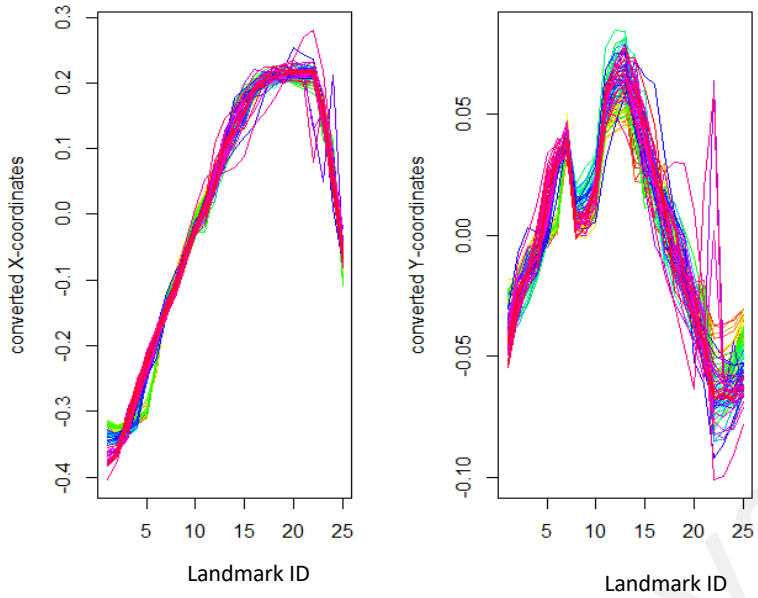
The univariate functional data is then represented as multivariate functional data, with n observations defined on d-dimensional domains using the *multiFunData* function (Happ-Kurz, 2020).

(a)



(b)

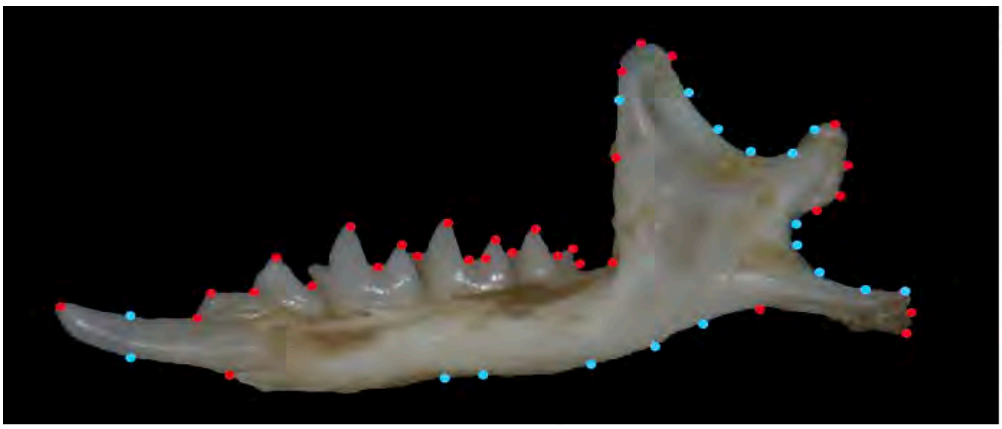
Figure 4.2: (a) 25 landmarks included for dorsal view of *C. malayana*. Landmarks and semilandmarks are represented by red and light blue dots, respectively. (b) 2D representation of the x and y-coordinates for the 25 landmarks of crania for the dorsal view; (c) 2D domains of converted functional data of landmark data for the dorsal view using FDGM method (specimens are represented by coloured lines) for: (i) Dimension 1 and (ii) Dimension 2



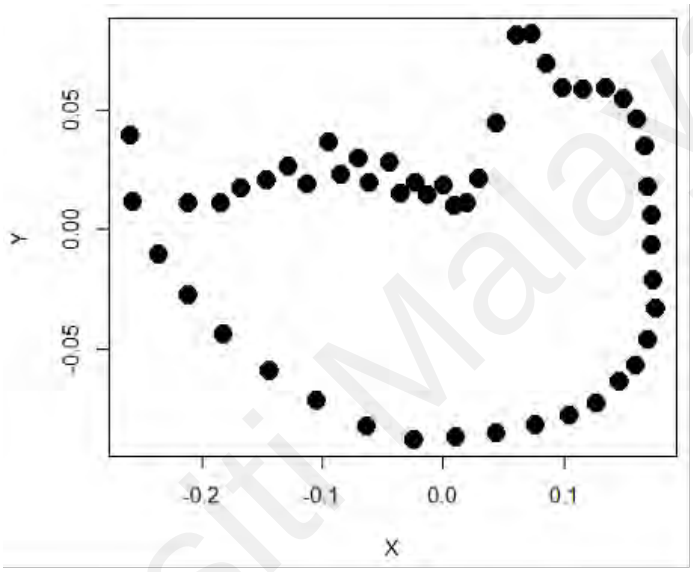
(i) (ii)

(c)

Figure 4.2, continued.

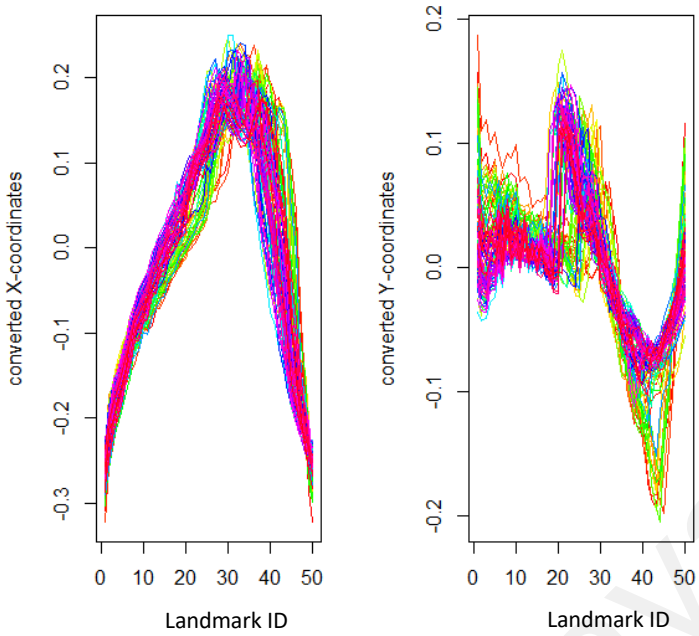


(a)



(b)

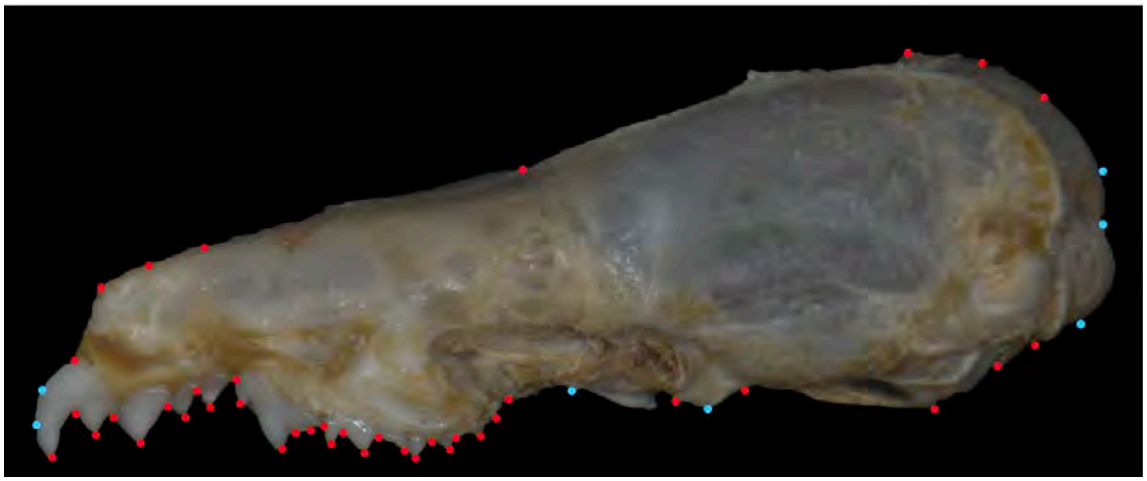
Figure 4.3: (a) 50 landmarks included for jaw view of *C. malayana*. Landmarks and semilandmarks are represented by red and light blue dots, respectively (b) 2D representation of the x and y-coordinates for the 50 landmarks of crania for the jaw; (c) 2D domains of converted functional data of the landmark data for the jaw view using the FDGM method (specimens are represented by coloured lines) for: (i) Dimension 1 and (ii) Dimension 2.



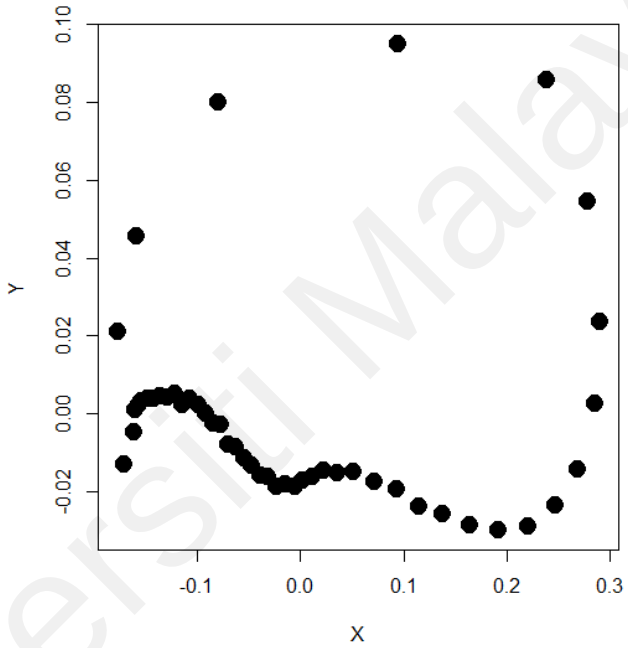
(i) (ii)

(c)

Figure 4.3, continued.

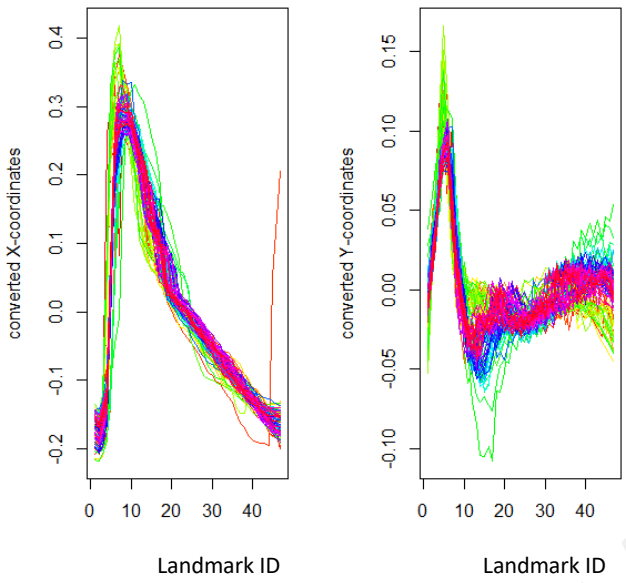


(a)



(b)

Figure 4.4: (a) 47 landmarks included for lateral view of *C. malayana*. Landmarks and semilandmarks are represented by red and light blue dots, respectively (b) 2D representation of the x and y- coordinates for the 47 landmarks of crania for the lateral view; (c) 2D domains of converted functional data of the landmark data for the lateral view using the FDGM method (specimens are represented by coloured lines) for: (i) Dimension 1 and (ii) Dimension 2.



(i) (c) (ii)

Figure 4.4, continued.

4.3.2 Multivariate Functional Principal Component Analysis for Craniodental Views of Shrew Specimens

After acquiring the multivariate functional data, the *MFPCA* package is used to compute the MFPCA estimates on the multivariate functional data, based on their univariate counterparts (Happ & Greven, 2018). The *MFPCA* function calculates MFPCA based on the observations that are independently and identically distributed (multivariate functional data obtained from the landmarks). The PCA basis functions are estimated from the multivariate functional data, $\mathbf{X}_k(t)$ using univariate functional principal component analysis (*u*FPCA), which is the most common basis expansion on a 1 - dimensional domain (Happ-Kurz, 2020). These basis functions were then applied on n observations based on the PACE (PCA through conditional expectation) approach (Yao et al. 2005). *u*FPCA is calculated by smoothed covariance using the *refund* package (Happ-Kurz 2020). In MFPCA, vectors are no longer considered PCs but are replaced by functions.

Consider the vector-valued stochastic process $\mathbf{X} = (\mathrm{X}, Y, Z)^{\mathrm{T}}$, representing functional random variables associated with standardised landmark coordinates, x, y and $z-$ coordinates respectively. For $1 \leq p \leq P$ (in our case $P = 3$), let I_x be a compact set in $\mathbb{R}$, with finite (Lebesgue) measure and such that : $X: I_x \to \mathbb{R}$ belongs to $\mathcal{L}^2(I_x)$, the space of square integrable functions on I_x . $\left(I_y, \mathcal{L}^2(I_y)\right)$ and $(I_z, \mathcal{L}^2(I_z))$ is similarly defined. The $P-$Fold Cartesian product of I_x and I_y denoted by $\mathbf{I} \coloneqq I_x \times I_y \times I_z$. So, $\mathbf{X}$ is a stochastic process indexed by $\mathbf{t} \in \mathbf{I}$ and taking values in the $P-$Fold Cartesian product space $\mathcal{H} \coloneqq \mathcal{L}^2(I_x) \times \mathcal{L}^2\left(I_y\right) \times \mathcal{L}^2(I_z)$.

Let the inner product $\langle\langle \cdot, \cdot \rangle\rangle: \mathcal{H} \times \mathcal{H} \to \mathbb{R}$,

$$\langle\langle f, g \rangle\rangle \coloneqq \sum_{p \in \{x,y,z\}} \langle f_p, g_p \rangle = \sum_{p \in \{x,y,z\}} \int_{I_p} f_p(t_p) g_p(t_p) dt_p,$$

$$f = \left(f_x, f_y, f_z\right)^T, g = \left(g_x, g_y, g_z\right)^T \in \mathcal{H}.$$

Then, $\mathcal{H}$ is a Hilbert space with respect to the scalar product $\langle\langle \cdot, \cdot \rangle\rangle$ (see (Happ and Greven 2018)). $|||\cdot|||$ is denoted by the norm induced by $\langle\langle \cdot, \cdot \rangle\rangle$.

4.3.3 Multivariate Karhunen-Loève Representation

Assume that $\mathbb{E}[\mathbf{X}(\mathbf{t})] \coloneqq \left(\mathbb{E}[\mathrm{X}(t_x)], \mathbb{E}\left[\mathrm{Y}(t_y), \mathbb{E}[\mathrm{Y}(t_z)]\right]\right)^{\mathbf{T}} = \mathbf{0}, \forall \mathbf{t} = \left(t_x, t_y, t_z\right)^T \in \mathbf{I}$.

Let C denote the 3×3 matrix-valued covariance function which, for $\mathbf{s}, \mathbf{t} \in \mathbf{I}$, is defined as

$$C(\mathbf{s}, \mathbf{t}) = \mathbb{E}[X(\mathbf{s}) X(\mathbf{t})^{\mathrm{T}}]$$

where the (p, q)th of the matrix $C(\mathbf{s}, \mathbf{t})$, for $1 \leq p, q \leq P$, is the covariance function between the $p-$th and the $q-$th components $\mathbf{X}$:

$$C_{p,q}(s_p, t_q) = \mathbb{E}\left[X_p(s_p) X_q(t_q)\right] = \mathrm{Cov}\left(X_p(s_p), X_q(t_q)\right),$$

$$s_p \in I_p, t_q \in I_q, p, \in \{x, y\}$$

In particular, $C_{p,q}(\cdot,\cdot)$ belongs to $\mathcal{L}^2(I_p \times I_q)$. Let Γ: $\mathcal{H} \to \mathcal{H}$ be the covariance operator of $\mathbf{X}$ on the Hilbert space $\mathcal{H}$, where for $f \in \mathcal{H}$ and $\mathbf{t} \in \mathbf{I}$, the qth component of $\Gamma f(\mathbf{t})$ is given by

$$(\Gamma f)^{(q)}(t_q) := \left\langle \langle C_{\cdot,q}(\cdot, t_q) f(\cdot) \rangle \right\rangle = \sum_{p=1}^{P} \int_{I_p} C_{p,q}(s_p, t_q) f_p(s_p) ds_p ,$$

$$s_p \in I_p, t_q \in I_q, f \in \mathcal{H}.$$

By the theory of Hilbert-Schmidt operators, there exists a complete orthonormal basis $\{\phi_j, j = 1,2,\dots\} \subset \mathcal{H}$ and a sequence of real numbers $\lambda_1 \geq \lambda_2 \geq \dots \geq 0$ such that

$$\Gamma\phi_j = \lambda_j \phi_j \quad \text{and} \quad \lambda_l \to 0 \quad \text{as } j \to \infty.$$

The λ_j's are the eigenvalues of the covariance operator Γ and the ϕ_j's are the associated eigenfunctions. The multivariate version of the Karhunen-Loève's representation is:

$$\mathbf{X}(\mathbf{t}) = \sum_{j=1}^{\infty} \xi_j \phi_j(\mathbf{t}), \mathbf{t} \in \mathbf{I},$$

with zero mean random variables $\xi_j = \langle \langle \mathbf{X}, \phi_j \rangle \rangle$ and $\mathrm{Cov}(\xi_j, \xi_l) = \lambda_l 1_{\{j=l\}}$. Let $J \geq 1$ and assume that the first J eigenvalues are nonzero, i.e. $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_J \geq \lambda_{(J+1)}$. Up to a sign, the elements of the MFPCA basis are characterised by:

$$\phi_1 = \underset{\phi}{^{argmax}} \langle \langle \Gamma\phi, \phi \rangle \rangle \text{ such that } |||\phi||| = 1,$$

$$\phi_2 = \underset{\phi}{^{argmax}} \langle \langle \Gamma\phi, \phi \rangle \rangle \text{ such that } |||\phi||| = 1, \text{ and } \langle \langle \phi, \phi_1 \rangle \rangle = 0,$$

$$\vdots$$

$$\phi_{J+1} = \underset{\phi}{^{argmax}} \langle \langle \Gamma\phi, \phi \rangle \rangle \text{ such that } |||\phi||| = 1, \text{ and } \langle \langle \phi, \phi_l \rangle \rangle = 0, \forall l \leq J.$$

Then, the truncated Karhunen-Loève expansion of the process **X** is

$$\mathbf{X}_{[J]}(\mathbf{t}) = \sum_{j=1}^{J} \xi_j \phi_j(\mathbf{t}),\ \mathbf{t} \in \mathbf{I},\ J \geq 1\ ;$$

and the truncated Karhunen-Loève expansion of the components of **X** is

$$X_{p,[J_p]}(t_p) = \sum_{j=1}^{J_p} \psi_{p,j} \tilde{\phi}_{p,j}(t_p), \qquad t_p \in I_p,\ J_p \geq 1,\ p \in \{x, y\};$$

where $\{\tilde{\phi}_{p,j}, j = 1,2,\dots\}$ is the univariate FPCA basis associated to the covariance operator Γ_p of X_p and the scores are $\psi_{p,j} = \langle X_p, \tilde{\phi}_{p,j} \rangle$. Happ and Greven (2018) derived a direct relationship between the truncated representations (4.9) of the single elements X_p and the truncated representation (4.8) of the multivariate functional data **X**.

The principal component elements are in general, unknown and have to be estimated from a sample that are possibly observed on different sparse grid points. These elements are the eigenvalues $\{\lambda_j\}_{j\geq 1}$, the eigenfunctions $\{\phi_j\}_{j\geq 1}$ and the scores $\{\xi_j\}_{j\geq 1}$. Given a sample of n i.i.d observations $\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(n)}$ of **X**, the estimation procedure for MFPCA consists:

1. For each element X_p, estimate a univariate FPCA based on the observations $X_p^{(1)}, \dots, X_p^{(n)}$ by estimating the variance function $K_p(\cdot,\cdot)$ of X_p as follows:

$$\widehat{K}_p(s,t) = \frac{1}{n-1} \sum_{i=1}^{n} X_p^{(i)}(s)\, X_p^{(i)}(t).$$

This results in the estimated eigenfunctions $\hat{\tilde{\phi}}_{p,j}$, and scores $\psi_{p,j}$, $i = 1, \dots, n$, $j = 1, \dots, J_p$ for a given truncation integer J_p.

2. Define the matrix $\boldsymbol{\Xi} \in \mathbb{R}^{n \times J}$ with $J = \sum_{p\in\{x,y,z\}} J_p$, where each row $(\psi_{1,1}^{(i)}, \dots, \psi_{1,J_x}^{(i)}, \dots, \psi_{3,1}^{(i)}, \dots, \psi_{3,J_z}^{(i)})$ contains the estimated scores for the 3 components of the i-th observation. Let's consider that the matrix $\mathbf{Z} \in \mathbb{R}^{J \times J}$ consisting of blocks $\mathbf{Z}^{(pq)} \in \mathbb{R}^{J_p \times J_q}$ with entries

$$Z_{jk}^{(pq)} = \mathrm{Cov}(\psi_{p,j}, \psi_{q,k}), \qquad j = 1, \dots, J_p,\ k = 1, \dots, J_q,\ p, q = 1,2,3.$$

An estimate $\hat{\mathbf{Z}} \in \mathbb{R}^{J\times J}$ of the matrix $\mathbf{Z}$ is given by

$$\hat{\mathbf{Z}} = \frac{1}{n-1}\boldsymbol{\Xi}^T\boldsymbol{\Xi}.$$

3. Perform a matrix eigen-analysis for $\hat{\mathbf{Z}}$ resulting in eigenvalues $\hat{\lambda}_j$ and construct the orthonormal eigenvectors $\hat{\mathbf{v}}_\mathbf{j}$.

4. Elements of the estimated multivariate eigenfunctions are given by

$$\hat{\phi}_{p,j}(t_p) = \sum_{k=1}^{J_p}\left[\hat{\mathbf{v}}_\mathbf{j}\right]_{p,k}\hat{\tilde{\phi}}_{p,k}(t_p), \quad t_p \in I_p\,, \;\; j = 1,\dots,J, p \in \{x,y,z\};$$

And the corresponding multivariate scores are calculated via

$$\hat{\xi}_\mathrm{j}^{(i)} = \sum_{p=1}^{P}\sum_{k=1}^{J_p}\left[\hat{\mathbf{v}}_\mathbf{j}\right]_{p,k}\psi_{p,k}^{(i)} = \boldsymbol{\Xi}_i.\hat{\mathbf{v}}_\mathbf{j}.$$

These estimated eigen values and functions are derived under the assumption of a finite sample size n and a finite Karhunen-Loève representation for each univariate function X_p.

4.3.4 Functional Linear Discriminant Analysis for Craniodental Views of Shrew Specimens

The MFPC scores from the landmarks were then applied in LDA to distinguish among the categories studied and the results were compared with the PCA of the GM approach. In terms of object recognition, it is generally believed that LDA tends to be superior compared to PCA (Martinez & Kak, 2001). LDA is a dimension reduction technique that is often used to model differences in groups. Functional linear discriminant analysis (FLDA) is an extension of linear discriminant analysis (LDA) to the case where the predictor variables are curves or functions (James & Hastie, 2001) of linear discriminant analysis (LDA) to the case where the predictor variables are curves or functions (James & Hastie, 2001).

FLDA enables the generation of classifications for new curves, offers an estimation of the discriminant function distinguishing between classes, and furnishes a one- or two-dimensional graphical depiction of a collection of curves (James & Hastie, 2001). The

number of PC scores used in LDA are obtained based on a threshold of 90% variation explained to compare rates of classification for both GM and FDGM methods. FLDA uses a spline curve, which is parameterised using a basis function multiplied by a d-dimensional coefficient vector to effectively transform the data into a single d-dimensional space (James & Hastie, 2001). This classifier also includes the random error to model observations from each individual (James & Hastie, 2001). The coefficient vector is then modelled using a Gaussian distribution with common covariance matrix for all classes by analogy with LDA (James & Hastie, 2001). The observed curves can then be pooled to estimate the covariance and mean for each class, which makes it possible to form accurate estimates for each individual curve based on only a few observations (James & Hastie, 2001).

Let M be the set of classes with Q denoted as the covariance matrix of the variables centered on the class mean, and B be predictions by the class means (Venables & Ripley, 2002). Let H be the $M \times W$ matrix of class means, where $W \geq 2$ represents the categorical variables. Denote G to be the $n \times M$ matrix of class indicator variables. Thus, the predictions are GH. $\bar{\rho}$ is the mean of the PC scores over the whole sample. The sample covariance matrices are as follows.

$$W = \frac{(\rho - G)^T(\rho - G)}{n - M}, B = \frac{(G - 1\bar{\rho})^T(G - \bar{\rho})}{M - 1},$$

where ρ are the selected PC scores.

LDA maximises the ratio of the separation of the class means to the within-class variance by maximising the ratio $\frac{a^T B a}{a^T W a}$ where a is the eigenvector of B corresponding to the largest eigenvalue (Fisher, 1936).

4.4. Classification Models

Machine learning has been extensively used in morphometric studies for classification and identification tasks (Tan et al., 2018). NB, SVM, and RF were chosen as classification models as these models were commonly used in many classifications related studies. Van der Plaat et al. (2021) applied NB and RF classifiers for species classification in plant genetic resources collections. GLM was one of the chosen classifiers to observe species distribution data at three fine scales: fine (Catalonia), intermediate (Portugal) and coarse (Europe) (Thuiller et al., 2003). The performances of the NB, SVM, RF and GLM methods on classification of species among the shrews were assessed using the principal component scores from functional data (MFPCA) and classical PCA scores. This was done using the *e1071, MASS and caret* packages in R. The combined analysis of all three views and each separate view was performed. Monte Carlo simulation was performed with 20 iterations to observe the possible output of each model. A brief description of these classification models is provided as follows:

i) Naïve Bayes

The naïve Bayes (NB) classification model is a classifier used to estimate the posterior probability to provide a mechanism that utilises predictors of the training data (Sammut & Webb, 2010). This approach has been successfully applied to species identification tasks, particularly when dealing with categorical or discrete features describing species characteristics. Based on the MFPC scores obtained from this study, the Bayes theorem can be written as follows:

$$P\left(c_i \middle| \hat{\xi}_1^{(i)}, \hat{\xi}_2^{(i)}, \hat{\xi}_3^{(i)}\right) = \frac{P(c_i)P\left(\hat{\xi}_1^{(i)}, \hat{\xi}_2^{(i)}, \hat{\xi}_3^{(i)} \middle| c_i\right)}{P\left(\hat{\xi}_1^{(i)}, \hat{\xi}_2^{(i)}, \hat{\xi}_3^{(i)}\right)} ,$$

where $\hat{\xi}_1^{(i)}, \hat{\xi}_2^{(i)}, \hat{\xi}_3^{(i)}$ represents the selected MFPC scores and c_i represents the three shrew species (*C. malayana, C. monticola and S. murinus*).

ii) Support Vector Machine

Support vector machine (SVM) addresses a multi-class problem as a single "all-together" optimisation. This classifier can be used to find a hyperplane in a 2-dimensional space that will separate the scores to their potential species. As this study emphasises on 2D, thus the equation of the hyperplane in the two domains can be given as follows:

$$y = \hat{\xi}_0^{(i)} + \hat{\xi}_1^{(i)} x_1 + \hat{\xi}_2^{(i)} x_2$$

$$= w_0 + \sum_{i=1}^{2} w_i x_i$$

$$= w_0 + w^T X$$

$$= b + w^T X$$

The three main hyperparameters in SVM are the cost parameter (C), gamma (γ) and kernel. The cost (C) is the penalty parameter of the error term which controls the trade-off between achieving a low training error and a low testing error. The gamma (γ) hyperparameter defines the influence of individual training samples and the kernel is used for mapping the input data into a higher-dimensional space. The radial basis function (RBF) is selected as the kernel function in this study due to its strong classification approach and its versatility in application without requiring prior knowledge of the dataset (Mustaqeem & Saqib, 2021). SVM-RBF can be defined as follows:

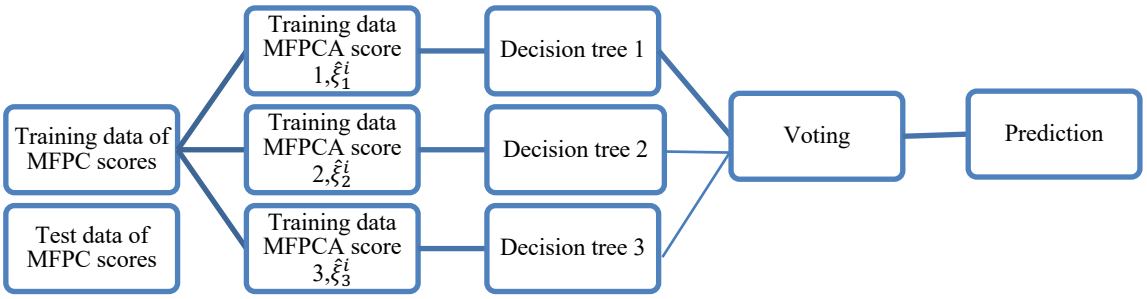
$$k(x_1, x_2) = \exp(-\gamma \left|| x_1 - x_2 \right||^2),$$

where $\gamma > 0, \gamma = \frac{1}{2\sigma^2}$.

iii) Random Forest

Random forests (RF) is an algorithm for classification developed by Breiman (2001) that is based on bootstrap aggregating or bagging that combines the predictions of multiple decision trees to make a final prediction. This helps to reduce the variance of

the individual trees, therefore reducing the overall expected prediction error of the random forest. The working algorithm of the RF classifier is as follows:



iv) Generalised Linear Model: Elastic Net Regularisation

The GLM classifier here is based on the elastic net penalty, which combines both L1 (LASSO) and L2 (ridge) penalties. In the context of geometric morphometrics, elastic net regularisation can be applied to GLMs to control the complexity of the model and prevent overfitting when analysing shape data. The alpha ($\propto$) is the parameter which controls the balance between L1 and L2 regularisation. Lambda (λ)is the penalty parameter that controls the strength of regularisation. This classifier based on the MFPC scores can be represented as $\eta_i = \beta_0 + \beta_1\hat{\xi}_1^{(i)} + \ldots + \beta_i\hat{\xi}_3^{(i)}$ with a link function (softmax function for multi-class classification) that describes how the mean, $\mathrm{E}(Y_i) = \mu_i$ depends on the linear predictor, $g(\mu_i) = \eta_i$. The GLM classifier also has a variance function that describes how the variance, $var(Y_i)$ depends on the mean, $var(Y_i) = \phi var(\mu_i)$ where the dispersion parameter, ϕ is a constant.

4.5 Results and Discussion

MFPCA using the functional data of all views combined gave a total of 31 eigenvalues. The first two MFPCs accounted for 81.56% of the total variation in the species of shrews. PCA using the GM method yields 89 principal components where the first two PCs explained 62.94%. The functional principal components show a comparable separation (Figure 4.5(b)) to the classical GM approach. *Suncus murinus* is shown to be well

separated in both methods (Figure 4.5 (a)(b)). Based on the principal component loadings, LM1 of dorsal view (i.e., the anterior most point of suture) is positively correlated to all three PCs indicating the strongest association to PC3. Thus, employing the FDGM approach has potential in examining the species variation of the shrews. When PCA is separately conducted on each view, the dorsal view gives the best separation for the three shrew species compared to the other two views for both GM and FDGM methods (Figure 4.6(a) and Figure 4.7(a)).

The dorsal view yielded a total of 10 MFPCs and the first two MFPCs explained 86.4% of the variation among the species. The GM method yields 46 PCs and the first two explained 59.24% of variation. The predicted MFPCA results gave a better separation among the three shrew species compared to the GM method.

There are 11 MFPCs for the jaw view where the first two MFPCs explained 89.31% of the variation in the species. There is a total of 89 classical PCs for the jaw view where the first two explained 73.13% of the variation. As for the lateral view, there is a total of 10 MFPCs and the total variation in species explained by the first two MFPCs is 90.90%. Out of the 89 PCs, the first 2 PCs of the GM approach for the lateral view explained 74.29% of total variation. Although *S. murinus* is somewhat separated, the jaw view and lateral view show poor separation for all three species for the GM approach (Figure 4.6(b) and (c)).

A comparable result for species separation can be observed in the FDGM approach (Figure 4.7(b) and (c)) for both views. The performance of the classification models based on individual craniodental views and the combination of all three is evaluated using the selected PC scores of both the FDGM and GM approaches as the PCs of all the craniodental views lie within the general rule of thumb threshold of 90% in the FDGM approach. The overall improvement in results for all the classification models when the FDGM approach is applied compared to the GM method is shown in Table 4.1.

The selected PC scores from GM and FDGM were then used in LDA to observe the percentage of separation among the three shrew species based on the craniodental views.

Based on GM, the percentage of separations achieved by the first discriminant function is 92.90%, second is 7.10% when all three craniodental views are combined. It is noticeable that the groups are quite well separated with FDGM showing better separation among the three species (Figure 4.8 (b)). The percentage of separations achieved by the first discriminant function in FDGM is higher compared to GM, which is 99.89 %.

Based on the results obtained in FLDA when the three craniodental views are observed separately, the dorsal view showed a distinct separation of *S. murinus* compared to the other two shrew species, which overlapped (Figure 4.10 (a)). This result is expected because *C. monticola* and *C. malayana* belong to the same genus. Besides that, both species inhabit similar ecological niches as insectivorous mammals, primarily found in forested habitats. Thus, the dorsal view of shrews plays an important role in capturing specific anatomical features of the shrews and providing unique insights into the overall shape and structure of the skull. This view provides a clear view of cranial sutures and landmarks, which are important for shrew species identification and comparative anatomy. Based on GM, the percentage of separations achieved by the first discriminant function is 92.90%, 98.00%, and 87.50% for dorsal, jaw and lateral respectively. The percentages of separation by the first discriminant function showed improvement in the FDGM method, which is 99.91% for the dorsal and jaw view, and 97.20% for the lateral view. *C. monticola* seems to be well grouped using the FDGM method for all views (Figure 4.10) compared to the GM method (Figure 4.9). In this thesis, the principal components utilised in LDA, FLDA, and other classification methods are derived from each craniodental views that collectively account for 90% of the explained variance. For the dorsal view, the first 3 MFPCs and the first 9 PCs are used for the FDGM and GM methods, respectively. For the jaw view, the first 2 MFPCs and the first 7 PCs are used

for the FDGM and GM methods, respectively. Similarly, for the lateral view, the first 2 MFPCs and the first 7 PCs are used for the FDGM and GM methods, respectively. When combining craniodental views, the first 4 MFPCs and the first 15 PCs are used for the FDGM and GM methods, respectively. FDGM needs fewer components than GM to account for the 90% explained variance threshold. FLDA leverages the full structure of functional data by considering the entire curve or shape as a single entity. This allows it to capture important patterns and relationships that might be missed if the data were simply reduced to a set of discrete measurements. In contrast, traditional LDA treats each measurement independently, potentially losing valuable contextual information. By modeling the data as functions, FLDA can better discriminate between classes based on the overall shape and structure of the data. This can lead to improved classification performance, especially in cases where the differences between classes are more subtle and spread across the entire function rather than concentrated in specific measurements. This is because FDGM represents shape variation as a continuous function over the entire curve or surface, whereas traditional GM typically represents shape using discrete landmark coordinates. This difference allows FDGM to capture more nuanced and continuous patterns of shape variation, which may be particularly beneficial for capturing subtle differences in shape between individuals or groups. FDGM also incorporates smoothing techniques or noise reduction algorithms as part of the functional data analysis process. This can help mitigate the effects of measurement error or noise in the shape data, leading to more distinct and well-defined groupings compared to the raw landmark data used in traditional GM methods. As the shape data represents functional curves or surfaces, FDGM explicitly models the functional dynamics of shape variation. This allows FDGM to capture temporal or spatial patterns of shape change, which may be critical for distinguishing between groups with subtle shape differences, such as those observed in *C. monticola*.

Distinct clusters of the shrew species are more prominent when the standardised landmarks of the three craniodental views combined are analysed by FDGM and GM methods. FDGM is a better solution as the outlines of the skulls are treated as continuous curves rather than discrete points (Ramsay & Silverman, 2005).

As shown in Figure 4.5, PCA based on GM does not give a better separation of the shrew species compared to MFPCA of the FDGM approach. When the three craniodental views were individually examined (Figure 4.6 and Figure 4.7), the dorsal view showed the clearest separation among the three shrew species using both approaches. This is because the dorsal view gives the most comprehensive view of the skull which includes landmarks from all the major cranial features. Based on the results obtained, this study reveals that the dorsal view of the shrew skulls can be the most informative view for distinguishing between the three shrew species.

The least favourable separations are observed for the jaw view (Figure 4.6 (b)). The MFPCA of the FDGM approach shows comparable results with that of GM's. As *C. monticola* and *C. malayana* belong to the same genus, there are similarities in the edges of the molar region for both species. The horseshoe effect present in the GM approach (Figure 4.6(b)) may indicate species turnover along environment gradients (Morton et al., 2017).

This effect has been commonly observed in ecological ordination obtained by PCA using the GM method (Podani & Miklos, 2002). The plots of the MFPCA scores (Figure 4.7(b)) reveal the presence of functional manifolds where the horseshoe effect is noticed (Wang et al., 2016). The lateral view also indicates an overlap between the two species. This is due to the similarity of the back curvature between the two as the region tends to be flat and a little sharp for *S. murinus*.

Considering that the FDGM framework relies on functions of craniodental curves

based on landmarks, the method shows viable results to the GM method in classification performance for all four models (Table 4.1). This is because MFPCA scores in machine learning can efficiently handle higher-dimensional data by capturing the functional nature, thus reducing dimensionality (Happ & Greven, 2018). Although PCs from GM reduce dimensionality, they might discard subtle but important variations by focusing on linear combinations of the original variables. Besides that, MFPCA provides scores that encapsulate smooth variations and inherent patterns in the data, making it easier for machine learning algorithms to discern meaningful distinctions between classes. The dorsal view gives the best rate of classification accuracy among the three views.

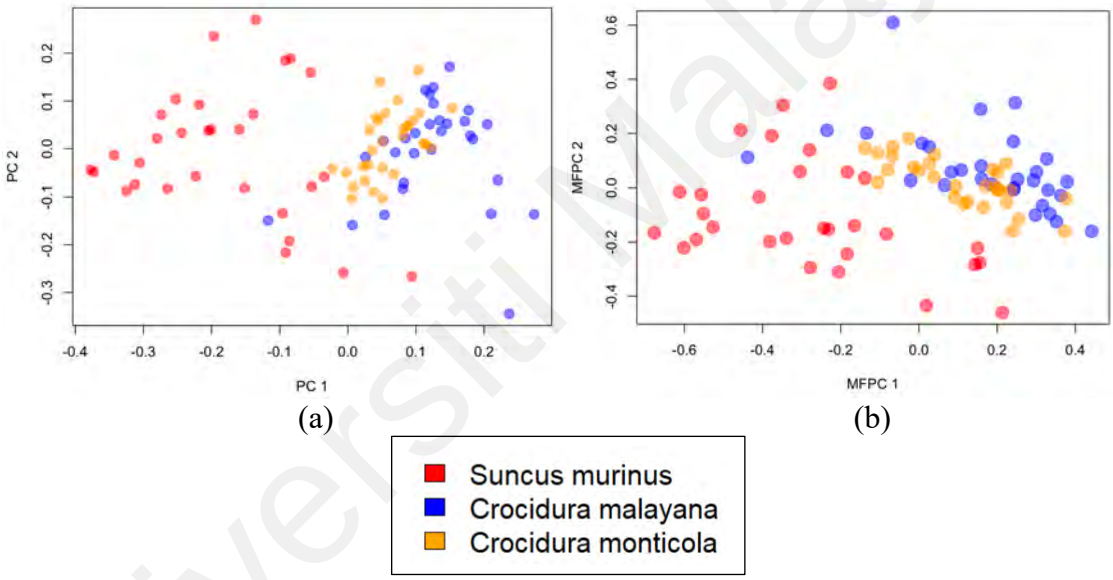


Figure 4.5: The PCs of the (a) GM (b) FDGM methods for all three views (dorsal, jaw and lateral combined)

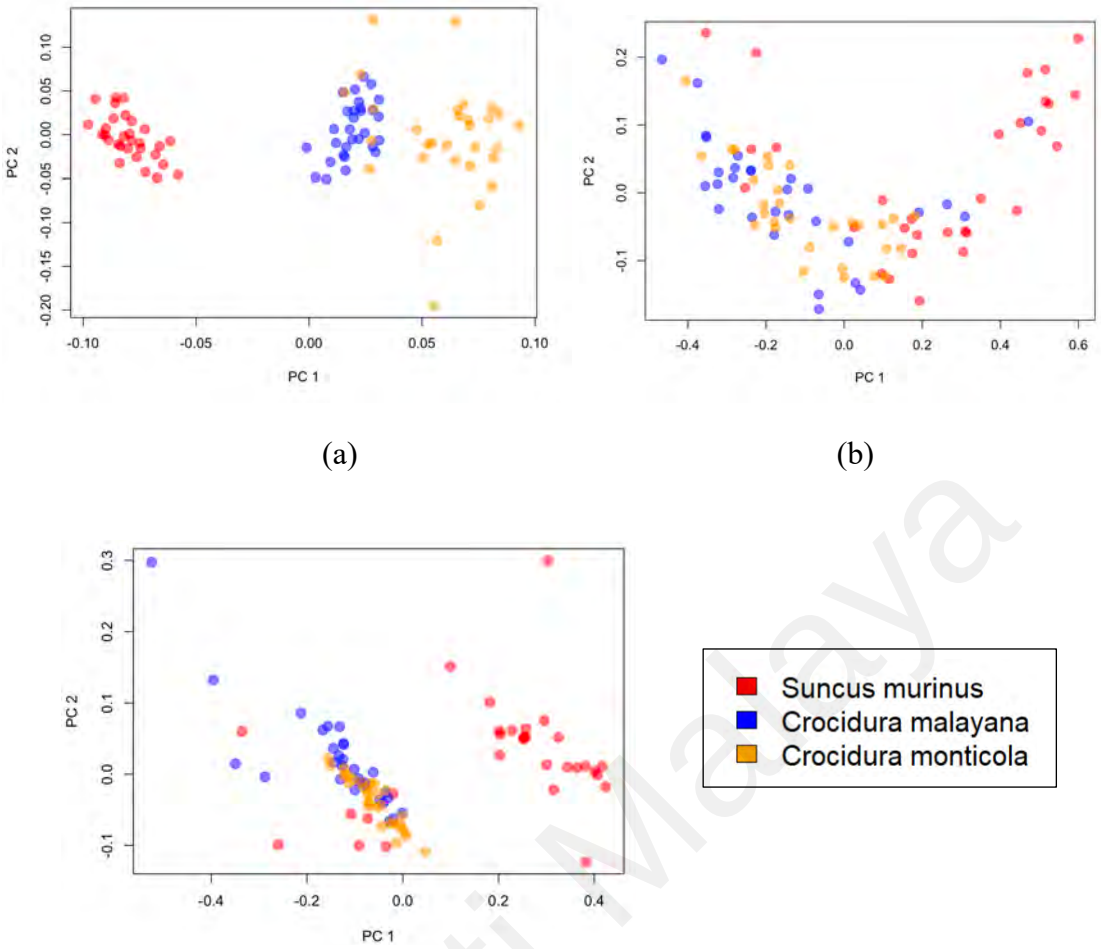


Figure 4.6: PCA plot using GM method for (a) dorsal view (b) jaw view (c) lateral view

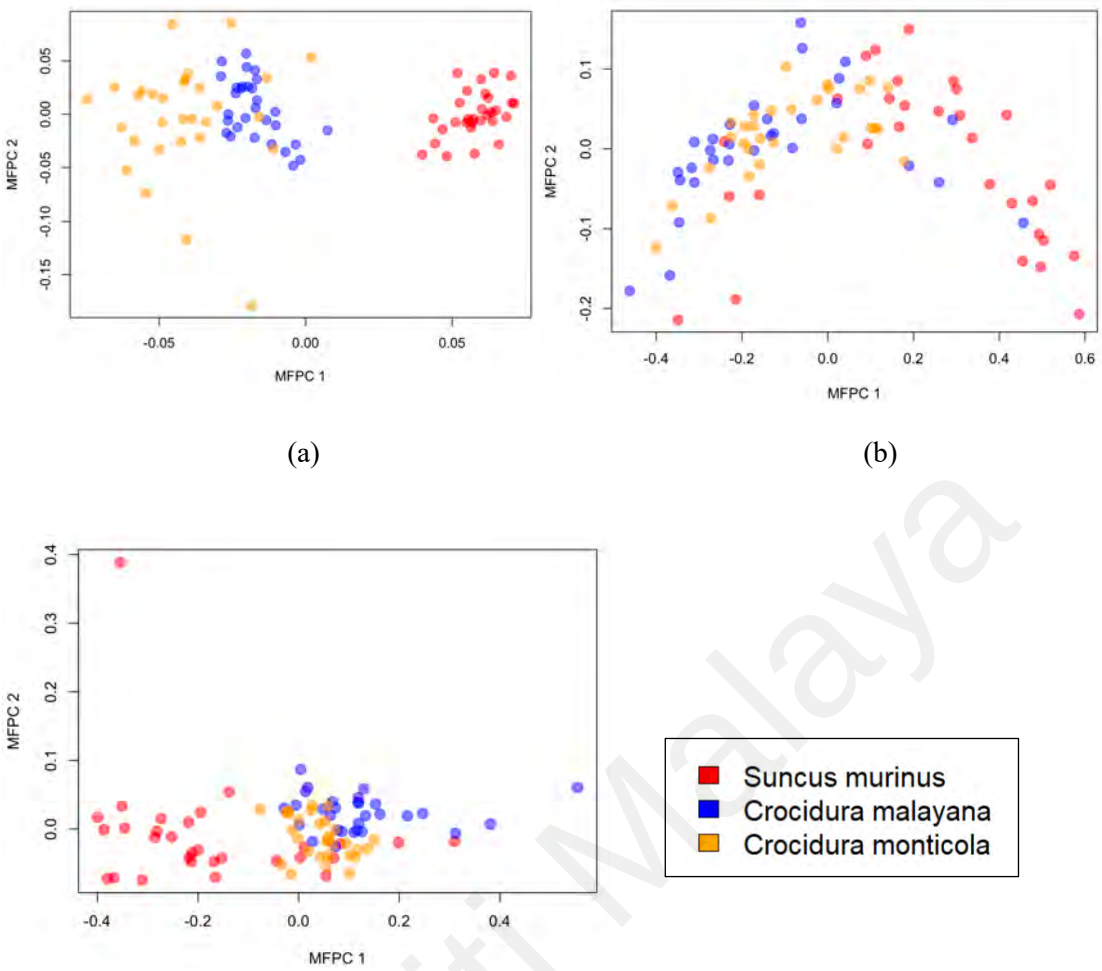


Figure 4.7: MFPCA plot using FDGM method for (a) dorsal view (b) jaw view (c) lateral view

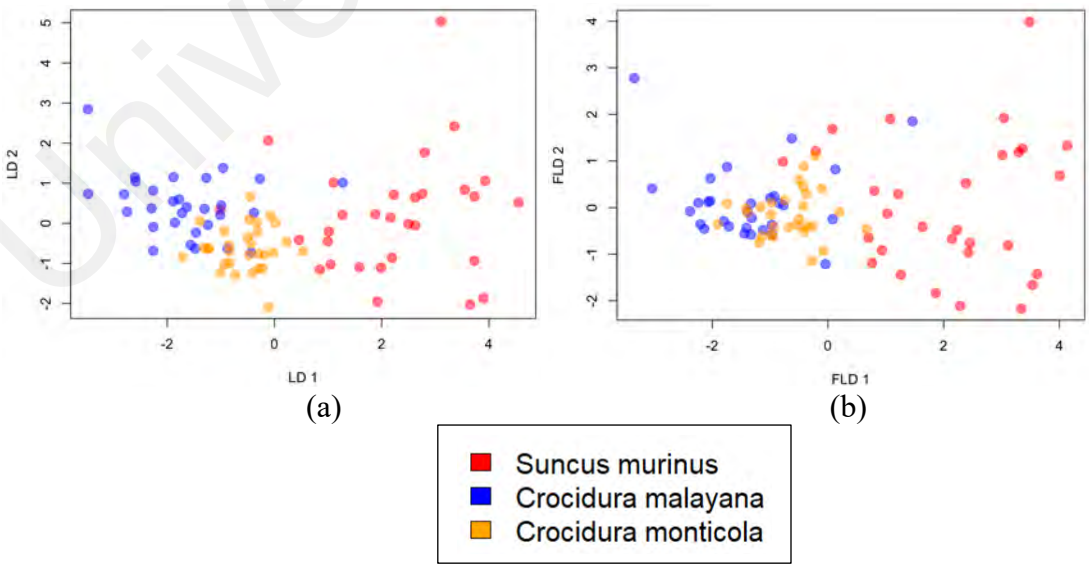


Figure 4.8: The LDs of the (a) GM (b) FDGM methods for all three views (dorsal, jaw and lateral combined)

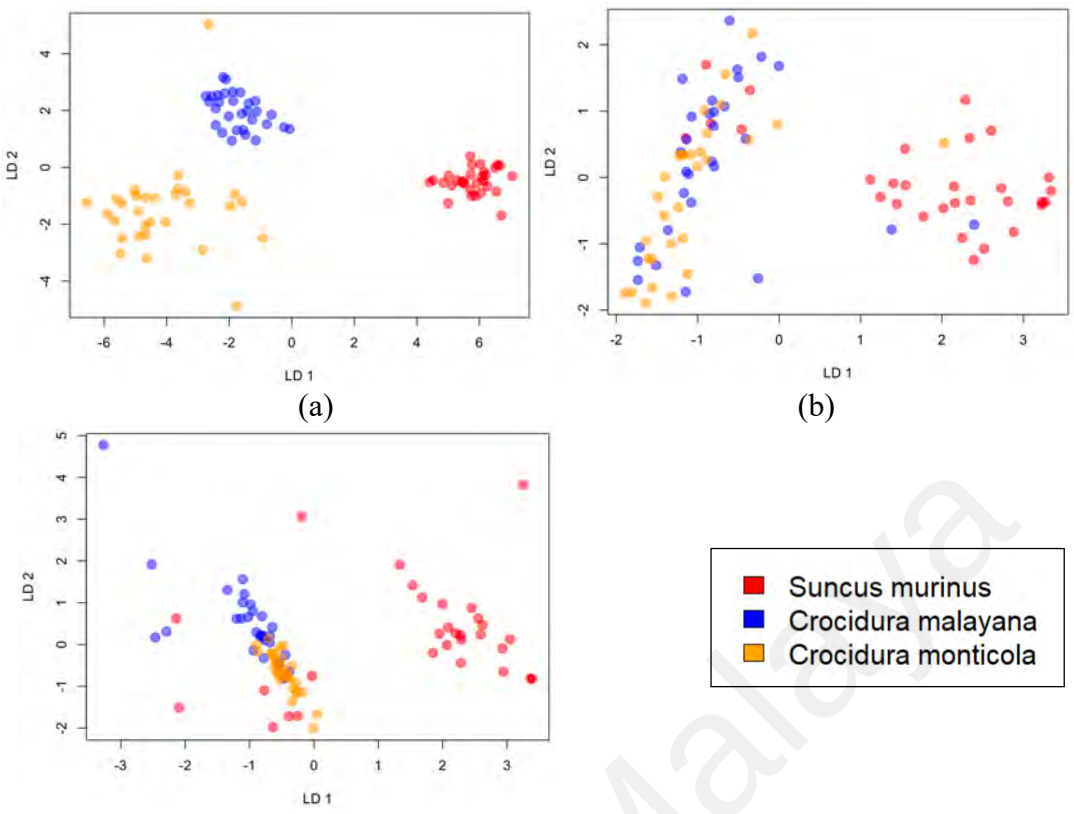


Figure 4.9: LDA plot using GM method for (a) dorsal view (b) jaw view (c) lateral view

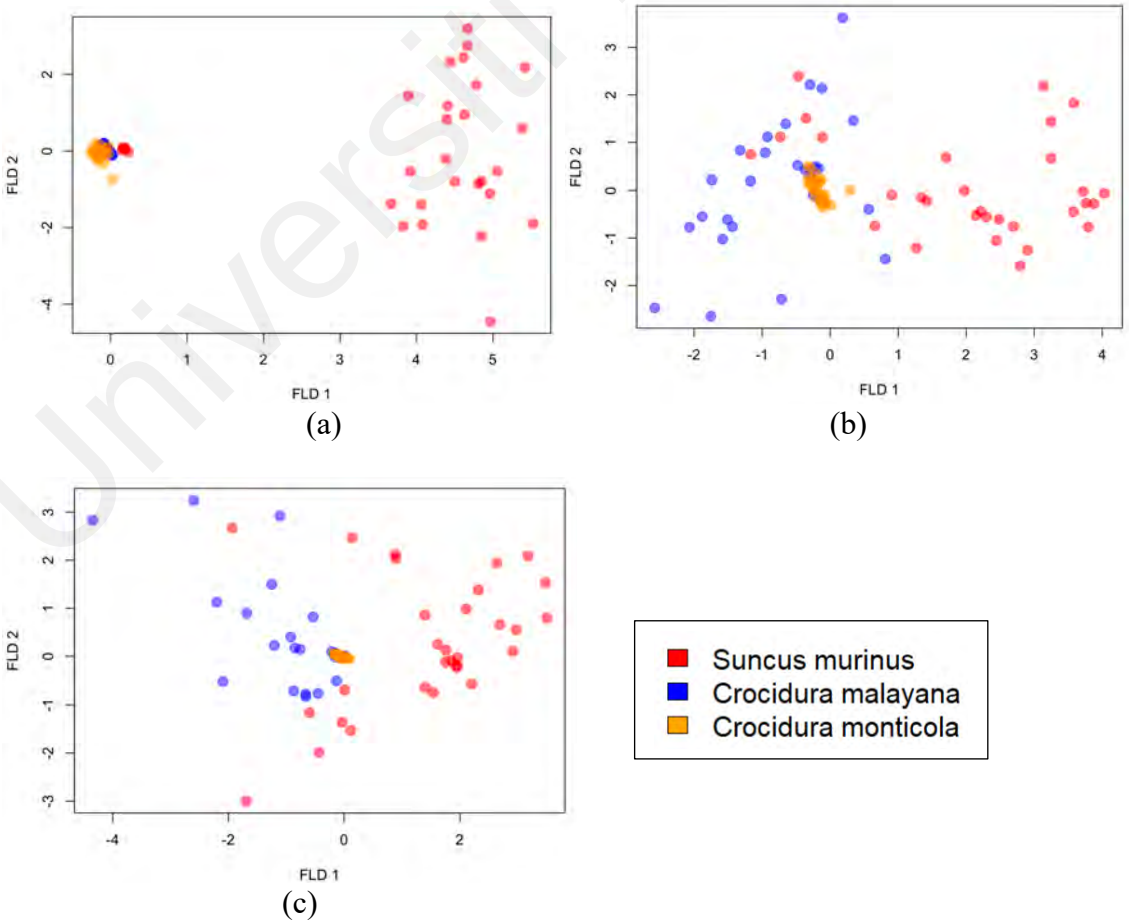


Figure 4.10: FLDA plot using FDGM method for (a) dorsal view (b) jaw view (c) lateral views

Table 4.1: The mean accuracy and the corresponding standard deviations (in brackets) on the test sample based on 20 replications using the FDGM and GM methods for views with dorsal, jaw and lateral combined. (b) individual views.

Classifiers	FDGM	GM
NB	0.976 (0.035)	0.881 (0.074)
SVM	0.962 (0.034)	0.962 (0.034)
RF	0.965 (0.025)	0.889 (0.084)
GLM	0.809 (0.057)	0.954 (0.034)
ANN	0.965 (0.035)	0.911 (0.050)

Table 4.2: The mean accuracy and the corresponding standard deviations (in brackets) on the test sample based on 20 replications using the FDGM and GM methods for individual craniodental views.

Classifiers	FDGM			GM		
	Dorsal	Jaw	Lateral	Dorsal	Jaw	Lateral
NB	0.969 (0.029)	0.565 (0.080)	0.820 (0.050)	0.993 (0.019)	0.578 (0.071)	0.841 (0.058)
SVM	0.950 (0.044)	0.557 (0.063)	0.800 (0.063)	0.950 (0.044)	0.557 (0.063)	0.800 (0.063)
RF	0.948 (0.044)	0.553 (0.092)	0.774 (0.072)	0.989 (0.022)	0.583 (0.098)	0.839 (0.053)
GLM	0.764 (0.096)	0.489 (0.055)	0.791 (0.066)	1.000 (0.000)	0.705 (0.078)	0.964 (0.038)
ANN	0.715 (0.144)	0.481 (0.031)	0.754 (0.077)	0.980 (0.028)	0.520 (0.085)	0.815 (0.079)

4.6 Simulation Studies For 2D Landmark Data

A simulation study is conducted to validate the general effectiveness of the methodology proposed in this work. The simulation was conducted using two approaches to assess the functional and classical PCAs. Method 1 simulates landmarks using the sim.coord function (Watanabe, 2018) where the coordinate data is generated with a specified number of specimens and landmarks from a multivariate normal distribution with zero mean and a variance-covariance structure using the mvrnorm function in the MASS R package (Venables & Ripley, 2002). Method 2 involves calculating the covariance matrix using the squared exponential function (Rasmussen, 2004). This method assumes that the coordinates are correlated with one another. The chosen PC scores of GM and FDGM were split into training data (70%) and test data (30%) to be applied into LDA and FLDA for both methods. The optimal number of iterations for both models is 100.

Model 1:

The simulation process where the coordinates are sampled from a multivariate normal distribution under a single variance-covariance scheme (unsmoothed data), which is based on the study conducted by Watanabe (2018) is as follows:

(1) Generate the 2-D landmark data, $\{(x_{k,\tau_1}, y_{k,\tau_1})^{\mathrm{T}}, \ldots, (x_{k,\tau_N}, y_{k,\tau_N})^{\mathrm{T}}\}$ for M groups, each with the same sample size with N landmarks per individual using the sim.coord function (Watanabe, 2018).

(2) Consider the PCA of GM and FPCA of FDGM based on 9. Calculate the cumulative proportion of variances explained for both methods for each iteration. Compute the means as well as standard errors for the 100 iterations.

(3) The approximate data is based on the dot product of the transpose of eigen vectors with transformed data for the PCA of GM and FPCA of FDGM. Calculate the reconstruction losses for both approaches. Compare the average and standard deviations of reconstruction loss for both methods for 100 iterations.

(4) Compare the LDA outputs for GM and FDGM.

Model 2:

The simulation process considering the mean and covariance functions Rasmussen (2004) implemented for 100 iterations involves the following steps:

(1) Generate sample points,$\{(x_{k,1}, y_{k,1})^{\mathrm{T}}, \ldots, (x_{k,N}, y_{k,N})^{\mathrm{T}}\}$, which are the test inputs used to define the mean and covariance functions based on N landmarks per individual and d dimensions.

(2) Generate a data frame which consists of normal random variates with zero mean and covariance sigma.

(3) Use the sample points obtained in Step 1 to calculate the covariance matrices based on the calculation done by (Rasmussen, 2004).

(4) Sample the function values, corresponding to the sample points from the joint posterior distribution by evaluating the mean and covariance matrix.

(5) Compute the covariance of the function values.

(6) Using the function values and covariance function obtained in Step 5, generate the 3-D landmark data for *M* groups, each with the same sample size with p landmarks per individual using the sim.coord.p function (Watanabe, 2018).

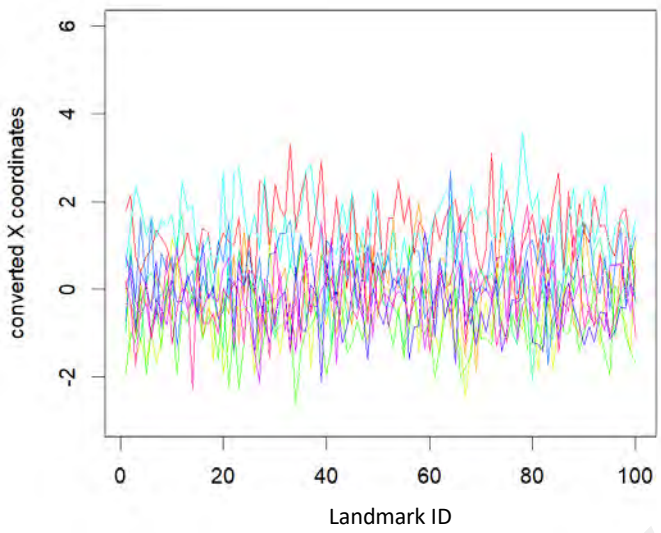
(7) The subsequent steps are similar to steps (2), (3) and (4) in Method 1.

An example of the comparison for the unsmoothed simulated landmark data, functional data, and the reconstructed functional data between the GM method and FDGM method (Model 1) is shown in Figure 4.11. The simulated data for Model 1 (Figure 4.11) is in an unsmoothed form and is not based on a functional data framework. Therefore, the results (Table 4.2) obtained are not favorable to the FDGM approach. For example, if the FDGM method assumes that the simulated data does not contain irregularities, this can lead to poor reconstruction. Therefore, this study uses Model 2, based on smoothed functional data which favours the functional data framework.

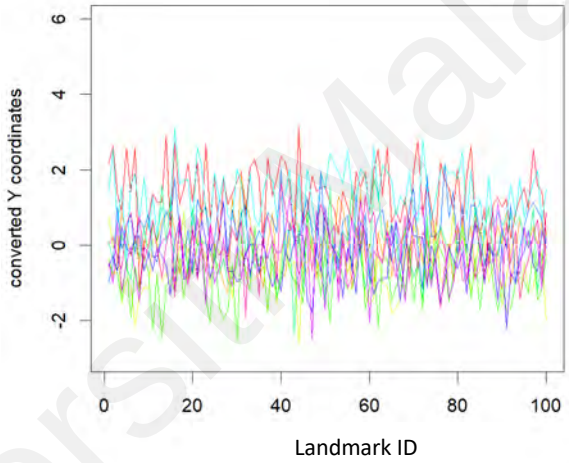
4.6.1 Results and Discussion of Simulation Studies

Table 4.2 shows that FDGM has a higher mean of cumulative variance for both simulation approaches used based on different numbers of groups and landmarks. For a fair comparison, the number of principal components used is based on a threshold value of 90% of variation explained. For example, for the first unsmoothed functional data simulation based on three groups for 20 landmarks, the number of PC for the GM method used is 27 and 1 MFPC using FDGM. As for the first smoothed functional data simulation based on three groups for 20 landmarks, the number of PC for the GM method are comparable with the FDGM method which is using three principal components. In terms of reconstructed data using for both models, FDGM has a lower error of reconstruction compared to GM. FDGM seems to obtain comparable classification rate based on the fLDA prediction results obtained in Table 4.3 using test data for the mean-covariance smoothed data. Incorporating machine learning algorithms into both models significantly improves Model 2 when using the FDGM method (Table 4.5). It can be observed that FDGM performs better in Model 2 because it more thoroughly considers the functional data framework. In contrast, Model 1 generates landmarks by assuming a multivariate normal distribution with a specified variance-covariance matrix. Model 2 applied Gaussian process regression (GPR) to generate the landmark data. The mean and

covariance are calculated through GPR to capture the smoothness in the data across a continuous domain. Therefore, while Model 1 is able to model correlated structures via correlation matrix, the data produced in Model 2 is aligned more closely with the principles of FDA due to the application of GPR. GM method outperforms FDGM in Model 1 as it takes more principal components to reach a threshold of 90% variation explained compared to FDGM. However, when considering both methods, MFPCA yields better results in the aspect of dimension reduction as it maximises variation explained with a reduced number of components compared to PCA. Based on the results, both NB and SVM outperform RF in terms of classification accuracy. This is because one of the practical implications of the RF model construction is that there is no way to replicate predictions without an actual forest. Future predictions thus require the original forest (including the original data) or a new forest that replicates the predictions with synthetic data (Prajwala, 2015). Model development is also more complex as each data set would generate a different model and there is no easy way to compare model parameters. Hence, validation of prediction models in separate population cohorts is likely to be challenging. NB can handle high-dimensional data well if the features are independent, leveraging its simplicity and the probabilistic approach whereas SVM can handle high-dimensional data effectively, especially with appropriate kernel functions that map data into higher-dimensional spaces. Besides that, for data including categorical variables with different number of levels, RFs are biased in favor of those attributes with more levels (Prajwala, 2015).

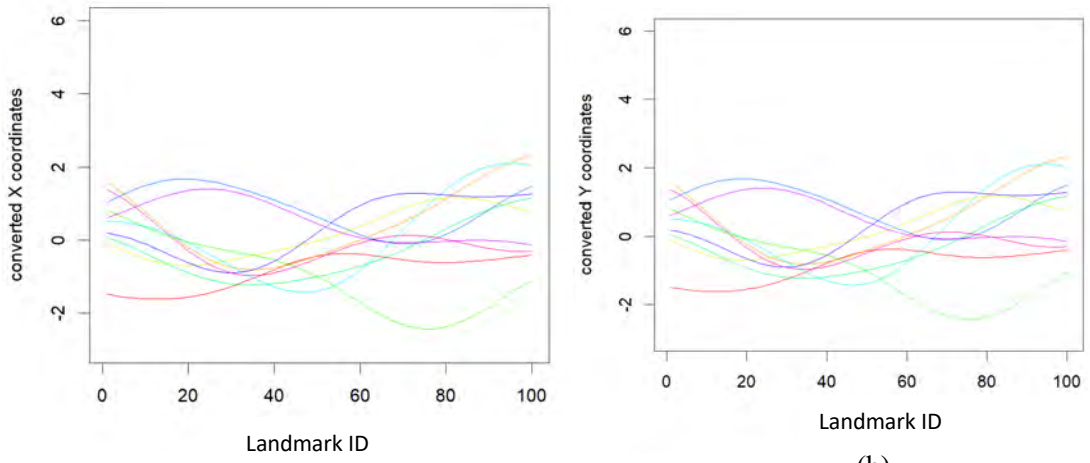


(a)



(b)

Figure 4.11: Comparison between functional data and reconstructed functional data based on Model 1 on 2D domains using the FDGM method (specimens are represented by coloured lines): (a) Dimension 1 (b) Dimension 2



(a) (b)

Figure 4.12: Comparison between functional data and reconstructed functional data based on Model 2 on 2D domains using the FDGM method (specimens are represented by coloured lines): (a) Dimension 1 (b) Dimension 2

Table 4.3: Mean (standard error values in parenthesis) of cumulative variance and error of reconstructed data for GM and FDGM methods for (i) Model 1 and (ii) Model 2 (100 simulations)

Number of groups	Number of landmarks	Model 1			
		Cumulative variance		Error of reconstructed data	
		GM	FDGM	GM	FDGM
3	20	0.946 (0.006)	0.957 (0.0061)	1.231 (0.115)	0.756 (0.056)
	50	0.947 (0.006)	0.954 (0.007)	1.229 (0.138)	0.757 (0.056)
	100	0.948 (0.007	0.954 (0.007)	1.238 (0.124)	0.766 (0.056)
4	20	0.946 (0.005)	0.957 (0.005)	1.383 (0.123)	0.761 (0.049)
	50	0.948 (0.004)	0.955 (0.006)	1.385 (0.114)	0.765 (0.051)
	100	0.948 (0.005)	0.953 (0.006)	1.373 (0.120)	0.767 (0.045)
5	20	0.945 (0.003)	0.956 (0.004)	1.485 (0.125)	0.761 (0.045)
	50	0.947 (0.005)	0.954 (0.006)	1.466 (0.123)	0.762 (0.041)
	100	0.948 (0.004)	0.953 (0.005)	1.464 (0.133)	0.773 (0.046)

(i)

Table 4.3, continued.

Number of groups	Number of landmarks	Model 2			
		Cumulative variance		Error of reconstructed data	
		GM	FDGM	GM	FDGM
3	20	0.945 (0.006)	0.956 (0.006)	1.226 (0.115)	0.115 (0.062)
	50	0.948 (0.006)	0.955 (0.007)	1.229 (0.122)	0.764 (0.054)
	100	0.948 (0.005)	0.953 (0.006)	1.224 (0.119)	0.761 (0.065)
4	20	0.945 (0.006)	0.955 (0.006)	1.386 (0.139)	0.753 (0.051)
	50	0.947 (0.006)	0.954 (0.006)	1.372 (0.143)	0.764 (0.049)
	100	0.948 (0.005)	0.954 (0.006)	1.386 (0.117)	0.768 (0.046)
5	20	0.944 (0.005)	0.955 (0.005)	1.475 (0.116)	0.761 (0.043)
	50	0.948 (0.004)	0.954 (0.004)	1.460 (0.127)	0.768 (0.043)
	100	0.950 (0.121)	0.954 (0.048)	1.497 (0.005)	0.773 (0.005)

(ii)

Table 4.4: Mean of proportion of trace of LDA and fLDA of test data for GM and FDGM methods for Model 1 and Model 2 (100 simulations)

Number of groups	Number of landmarks	Model 1		Model 2	
		GM	FDGM	GM	FDGM
3	20	0.333 (0.047)	0.283 (0.070)	0.220 (0.038)	0.447 (0.077)
	50	0.316 (0.117)	0.416 (0.117)	0.206 (0.043)	0.353 (0.122)
	100	0.350 (0.023)	0.333 (0.000)	0.260 (0.072)	0.400 (0.047)
4	20	0.325 (0.070)	0.287 (0.053)	0.130 (0.040)	0.220 (0.033)
	50	0.375 (0.035)	0.337 (0.123)	0.240 (0.074)	0.300 (0.047)
	100	0.225 (0.035)	0.225 (0.000)	0.170 (0.041)	0.215 (0.051)
5	20	0.220 (0.028)	0.210 (0.070)	0.084 (0.038)	0.224 (0.055)
	50	0.300 (0.004)	0.240 (0.059)	0.089 (0.028)	0.285 (0.071)
	100	0.180 (0.036)	0.160 (0.052)	0.077 (0.037)	0.214 (0.061)

Table 4.5: Mean of classification accuracy of classifiers for (i) GM and (ii) FDGM methods for Model 1 (100 simulations)

Number of groups	Number of landmarks	Model 1				
		GM				
		NB	SVM	RF	GLM	ANN
3	20	0.411 (0.015)	0.400 (0.125)	0.433 (0.047)	0.222 (0.031)	0.355 (0.031)
	50	0.422 (0.000)	0.355 (0.031)	0.355 (0.000)	0.211 (0.109)	0.411 (0.078)
	100	0.433 (0.204)	0.422 (0.062)	0.477 (0.109)	0.244 (0.031)	0.411 (0.078)
4	20	0.308 (0.035)	0.316 (0.000)	0.266 (0.070)	0.175 (0.011)	0.291 (0.058)
	50	0.383 (0.000)	0.316 (0.070)	0.325 (0.082)	0.250 (0.023)	0.325 (0.058)
	100	0.408 (0.011)	0.391 (0.035)	0.341 (0.011)	0.141 (0.035)	0.333 (0.000)
5	20	0.313 (0.028)	0.246 (0.009)	0.260 (0.028)	0.153 (0.009)	0.246 (0.103)
	50	0.280 (0.018)	0.246 (0.028)	0.286 (0.028)	0.193 (0.009)	0.206 (0.084)
	100	0.313 (0.028)	0.246 (0.009)	0.260 (0.028)	0.153 (0.009)	0.246 (0.103)

(i)

Table 4.5, continued.

Number of groups	Number of landmarks	Model 1				
		FDGM				
		NB	SVM	RF	GLM	ANN
3	20	0.333 (0.125)	0.366 (0.141)	0.322 (0.141)	0.177 (0.031)	0.333 (0.157)
	50	0.355 (0.062)	0.333 (0.031)	0.388 (0.109)	0.288 (0.031)	0.233 (0.078)
	100	0.322 (0.141)	0.322 (0.047)	0.311 (0.031)	0.244 (0.031)	0.211 (0.015)
4	20	0.266 (0.023)	0.308 (0.011)	0.241 (0.035)	0.225 (0.058)	0.208 (0.011)
	50	0.333 (0.000)	0.266 (0.047)	0.300 (0.000)	0.150 (0.023)	0.225 (0.035)
	100	0.291 (0.011)	0.308 (0.035)	0.241 (0.012)	0.175 (0.035)	0.191 (0.082)
5	20	0.260 (0.028)	0.220 (0.009)	0.213 (0.018)	0.146 (0.018)	0.160 (0.018)
	50	0.253 (0.000)	0.253 (0.056)	0.286 (0.028)	0.166 (0.009)	0.193 (0.009)
	100	0.260 (0.028)	0.220 (0.009)	0.213 (0.018)	0.146 (0.018)	0.160 (0.018)

(ii)

Table 4.6: Mean of classification accuracy of classifiers for (i) GM and (ii) FDGM methods for Model 2 (100 simulations)

Number of groups	Number of landmarks	Model 2				
		GM				
		NB	SVM	RF	GLM	ANN
3	20	0.500 (0.047)	0.235 (0.133)	0.027 (0.036)	0.467 (0.125)	0.613 (0.086)
	50	0.320 (0.144)	0.160 (0.059)	0.004 (0.059)	0.246 (0.086)	0.413 (0.119)
	100	0.387 (0.136)	0.200 (0.047)	0.013 (0.030)	0.493 (0.089)	0.426 (0.036)
4	20	0.360 (0.153)	0.133 (0.053)	0.000 (0.000)	0.373 (0.121)	0.347 (0.145)
	50	0.293 (0.101)	0.120 (0.056)	0.000 (0.000)	0.347 (0.110)	0.347 (0.159)
	100	0.347 (0.185)	0.147 (0.087)	0.000 (0.000)	0.467 (0.133)	0.280 (0.109)
5	20	0.253 (0.128)	0.067 (0.047)	0.000 (0.000)	0.307 (0.101)	0.360 (0.112)
	50	0.304 (0.127)	0.076 (0.090)	0.000 (0.000)	0.314 (0.074)	0.286 (0.050)
	100	0.276 (0.105)	0.095 (0.065)	0.000 (0.000)	0.247 (0.114)	0.295 (0.153)

(i)

Table 4.6, continued.

Number of groups	Number of landmarks	Model 2				
		FDGM				
		NB	SVM	RF	GLM	ANN
3	20	0.633 (0.237)	0.700 (0.047)	0.667 (0.149)	0.787 (0.247)	0.950 (0.960)
	50	0.800 (0.262)	0.680 (0.268)	0.586 (0.246)	0.586 (0.165)	0.840 (0.160)
	100	0.760 (0.138)	0.786 (0.314)	0.720 (0.272)	0.720 (0.207)	0.693 (0.252)
4	20	0.600 (0.287)	0.720 (0.172)	0.572 (0.121)	0.640 (0.180)	0.613 (0.231)
	50	0.546 (0.087)	0.653 (0.119)	0.706 (0.121)	0.680 (0.173)	0.773 (0.121)
	100	0.627 (0.161)	0.640 (0.269)	0.453 (0.159)	0.626 (0.238)	0.667 (0.282)
5	20	0.867 (0.094)	0.720 (0.166)	0.693 (0.293)	0.707 (0.101)	0.853 (0.172)
	50	0.876 (0.202)	0.667 (0.128)	0.828 (0.153)	0.752 (0.179)	0.771 (0.128)
	100	0.676 (0.194)	0.629 (0.246)	0.628 (0.285)	0.781 (0.120)	0.610 (0.156)

(ii)

4.7 Conclusion

In this chapter, the use of FDGM on landmark data is proposed to study the shapes of the dorsal, lateral, and jaw of shrew skulls in a functional form. The findings suggest that FDGM shows comparable results with GM for classification among the three species. The number of selected components in MFPCA can affect the classification quality. Therefore, a threshold of 90% explained variance is used to select the principal components for the GM and FDGM methods for fair comparison. Based on the results obtained, FDGM requires fewer components than GM to reach the 90% explained variance threshold. In addition, the results also revealed that the dorsal view emerges as the best representation for classifying the species in both approaches. The proposed approach utilises data smoothing to represent landmark coordinates as a function derived

from raw data, enhancing pattern clarity, and making it a potentially useful tool in morphometrics research. However, FDGM may encounter challenges in accurately capturing complex and non-linear shape transformations. This is because biological structures often exhibit complex shape transformations influenced by a myriad of factors, such as genetic variation, developmental processes, and environmental influences. Capturing these complex shape variations accurately with FDGM may require more sophisticated modeling techniques and larger, more diverse datasets. Additionally, integrating FDA techniques with GM requires careful data preprocessing and analytical methods to mitigate biases or errors. Despite these challenges, FDGM has the potential to analyse shape variation by modeling shape changes as continuous functions. This departure from traditional discrete landmark-based methods allows for a more comprehensive representation of shape, capturing subtle variations and non-linear transformations more effectively. By exploring the theoretical and practical advancements offered by FDGM, this study aims to contribute to the methodological toolkit of GM and facilitate more accurate and insightful analyses of biological shape data. Additionally, FDGM integrates principles from functional data analysis with GM, providing a more robust framework for analysing shape data. Practically, FDGM enhances the accuracy and sensitivity of shape analysis by enabling the examination of shape changes along continuous curves or surfaces.

This can lead to more precise identification of shape differences between groups and better understanding of shape variation within populations. Future studies can address these challenges and further explore the potential of FDGM. Additionally, ongoing research on three-dimensional FDGM extensions holds promise for further enhancing morphometrics analysis.

CHAPTER 5: FDGM IN 3D GEOMETRIC MORPHOMETRICS

5.1 Introduction

Craniodental morphology, the study of skull and dental structures, plays a pivotal role in unraveling the evolutionary biology, taxonomy, and ecological adaptations of marsupials. Marsupials, a diverse group of mammals primarily inhabiting Australia and the Americas (Beck et al., 2022), display a wide array of craniodental features reflecting their diverse diets. By scrutinising these features, researchers can glean insights into the evolutionary trajectories and adaptive strategies that have enabled marsupials to thrive in various environments. Morphological analyses often employ GM to discern subtle differences and similarities among species, shedding light on their evolutionary relationships and ecological roles.

Astúa et al. (2000) conducted a comprehensive analysis of cranial shape variation among six species representing the six largest living genera of the New World marsupial family *Didelphidae*. Utilising 2D landmark data, they captured and digitised video images of the skull and mandible for each species, providing a detailed exploration of cranial morphology within this taxonomic group. Their findings underscored the distinctiveness among species, emphasising the significant role of ecological factors in shaping cranial morphology (Astúa et al., 2000). Viacava et al. (2022) employed 3D GM of the cranium to enhance taxonomic differentiation and offer ecomorphological insights into a cryptic divergence within the carnivorous marsupial genus *Antechinus*. Their study highlighted the utility of 3D GM in elucidating the adaptive origins and potential threats to mammalian diversity, offering valuable perspectives for conservation planning in the face of environmental change (Viacava et al., 2022).

Butler et al. (2021) investigated the relationship between cranial and mandibular shape variation of extant and extinct macropodiforms, considering ecological factors such as diet, locomotion, and body mass. Utilising 3D GM analysis, they examined 42 living species and eight extinct species from two radiations, including the extinct clade of *Balbaridae* and early representatives of the extant *Macropodidae*. Their study revealed strong correlations between dietary class (fungivore, browser, grazer, mixed feeder) and cranial shape variation, along with significant associations between cranial shape and locomotor mode and body mass. These findings underscored the importance of integrating morphometric analyses with ecological and phylogenetic considerations to deepen our understanding of the feeding ecology and evolutionary history of extinct kangaroos and their adaptation to changing environments (Butler et al., 2021).

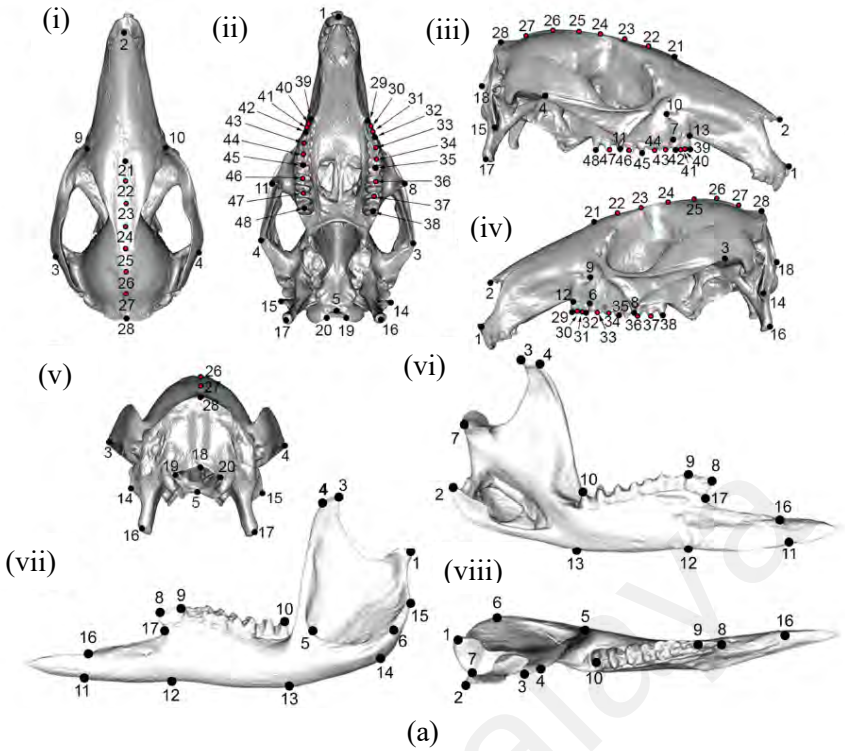
This thesis revolves around GM based on 3D landmarks data of kangaroos to investigate the relationship between variation in cranial and mandibular shape of extant macropodiformes with their dietary categories.

A multivariate functional principal component analysis (MFPCA) is then performed to produce interpretable descriptive analysis of the functional data obtained. The principal component (PC) scores obtained from both GM and FDGM are used to construct the linear discriminant analysis (LDA) model.

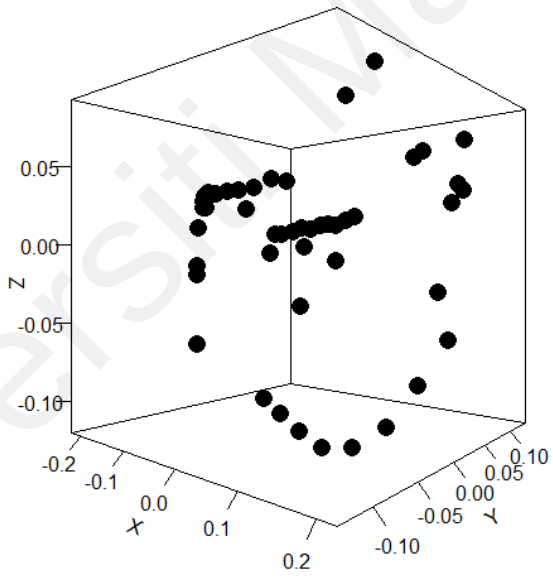
5.2 Functional Data Geometric Morphometrics in 3D Landmark Data

Landmark registration offers a straightforward approach that is used to detect and align some specific data points for each observation to the corresponding mean value, which provides a better representation of the mean in terms of amplitude variation. Each observation is vector-valued, as three spatial coordinates which are the x, y and $z-$ coordinates are involved. To implement functional data in an object-oriented way, the raw data is converted into functions.

These three univariate functional datasets composed a multivariate functional data, with n outlines, each yielding a vector of n observations defined as a $3-$dimensional functional domain. The *MFPCA* (Happ-Kurz, 2020) package is used to perform the conversion to functional data (Ramsay & Silverman, 2005). Since the three-dimensional landmark data is an extension of the two-dimensional case, where the individual data vector has been extended from length two to three, the methodology of analysis remains the same as in the two-dimensional case. After acquiring the multivariate functional data, MFPCA is performed using the univariate functional principal components. The PCA basis functions are estimated from the multivariate functional data.

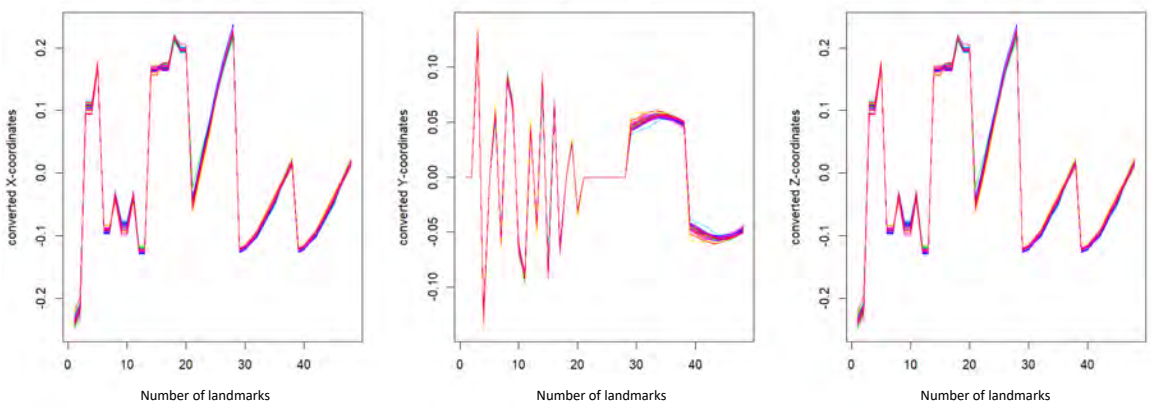


(a)



(b)

Figure 5.1: (a) 48 landmarks included for crania: 30 single landmarks and 18 semilandmarks in (i) dorsal view (ii) ventral view (iii) lateral right view (iv) lateral left view and (v) posterior view and for dentaries in (vi) lateral right view, (vii) lateral left view and (viii) occlusal view (Photo sourced from Butler et al., 2021). Single landmarks are represented by black dots while semilandmarks are represented by red dots with a black outline; (b) 3D representation of the x, y and $z-$ coordinates for the 48 symmetric shape landmark data of crania; (c) 3D domains of converted functional data of the symmetric shape landmark data using the FDGM method (specimens are represented by coloured lines) for: (i) Dimension 1, (ii) Dimension 2, (iii) Dimension 3



(i) (ii) (iii)

(c)

Figure 5.1, continued.

5.3 Simulation Studies for 3D Landmark Data

A simulation study for 3D landmarks is conducted to validate the general effectiveness of the methodology proposed in this work. Method 1 simulates landmarks using the sim.coord function (Watanabe, 2018) where the coordinate data is generated with a specified number of specimens and landmarks from a multivariate normal distribution with zero mean and a variance-covariance structure using the mvrnorm function in the MASS R package (Venables & Ripley, 2002). Method 2 involves calculating the covariance matrix using the squared exponential function (Rasmussen, 2004). This method assumes that the coordinates are correlated with one another. The selected PC scores of GM and FDGM were split into training data (70%) and test data (30%) to be applied into LDA and FLDA for both methods. The optimal number of iterations for both models is 100.

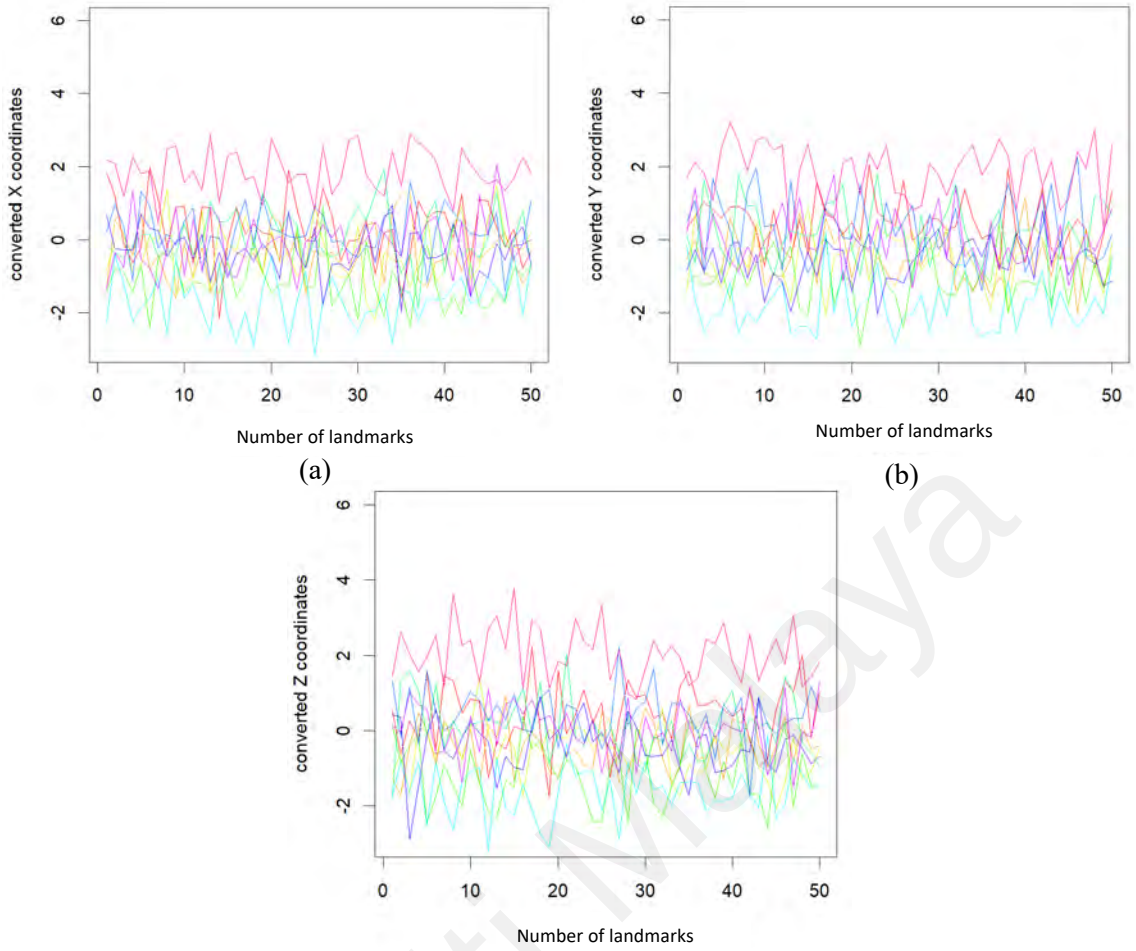


Figure 5.2: Functional data based on Model 1 on three dimensional domains using the FDGM method (specimens are represented by coloured lines): (a) Dimension 1 (b) Dimension 2 (c) Dimension 3

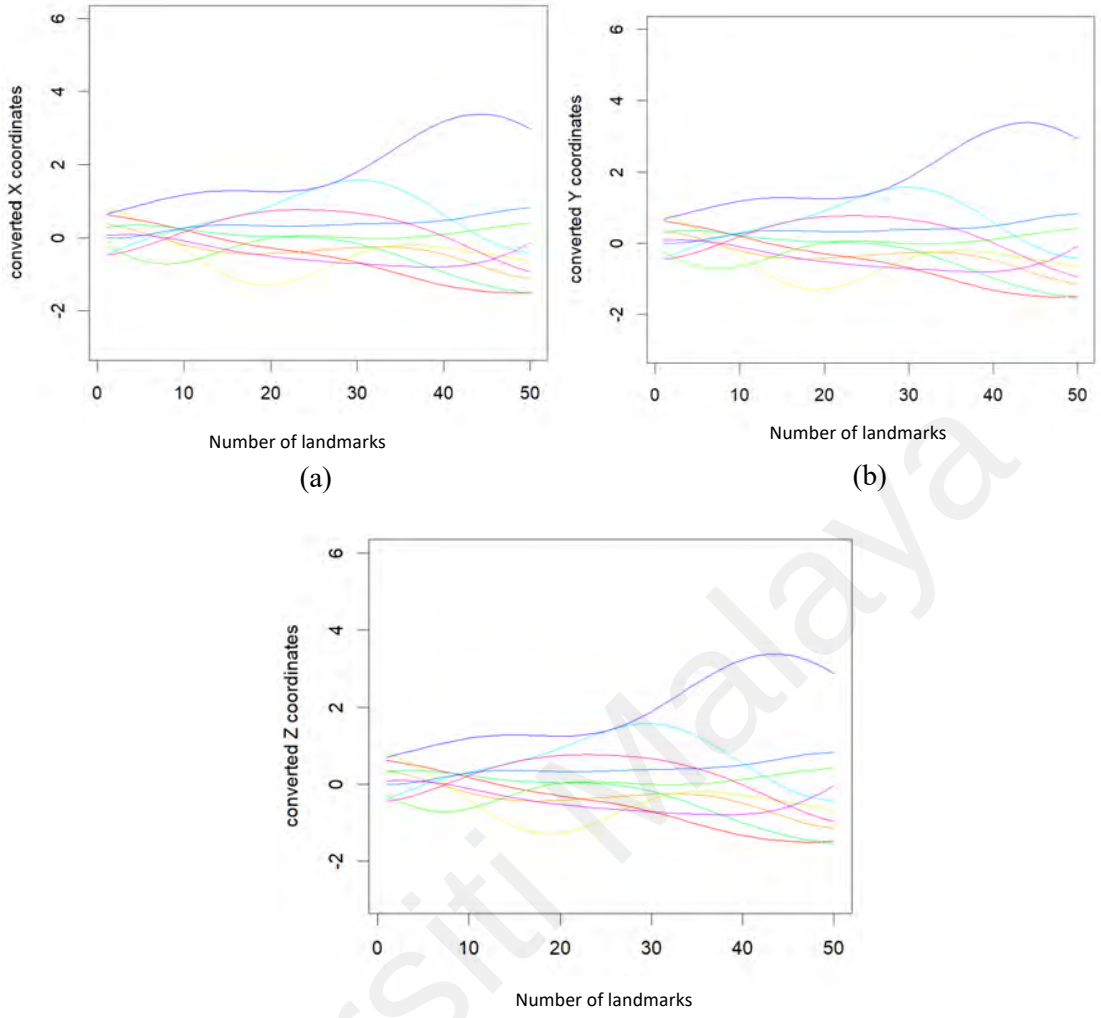


Figure 5.3: Functional data based on Model 2 on 3D domains using the FDGM method (specimens are represented by coloured lines): (a) Dimension 1 (b) Dimension 2 (c) Dimension 3

5.3.1 Results of Simulation Studies

Table 5.1 shows that FDGM has higher mean of cumulative variance for both simulation approaches used based on different numbers of groups and landmarks. In terms of reconstructed data using Model 1, GM performs better than FDGM due to the underlying simulation method which is biased towards multivariate normal coordinates that can easily be reconstructed to its original form based on the classical PCs. FDGM can be considered as comparable although some information may be lost when the points are converted into functions.

However, Model 2 favours FDGM in terms of the performance of MFPCA as well as the mean error of the reconstructed data, with the PCA of GM having larger errors. Table 5.2 shows that FDGM has a higher mean of proportion of trace using training data for the mean-covariance smoothed data based on different number of groups and landmarks. In terms of unsmoothed data (with weak correlation), GM seems to perform better than FDGM, similar to the cumulative proportion of variance. This may be due to unsmoothed landmark data construction in Model 1 without any functional structure, which means that it does not consider any patterns in the data. Model 2, on the other hand, includes the functional data structure. This means that it considers the patterns in the data, which makes it more likely to be accurate than Model 1, as it considers more information about the data. FDGM seems to significantly improve the classification rate based on the FLDA prediction results obtained in Table 5.3 using test data for the mean-covariance smoothed data. The results in Table 5.3 give the correct percentage of classification that is compared with the training data.

Table 5.1: Mean (standard error values in parenthesis) of cumulative variance and error of reconstructed data for GM and FDGM methods for the entire (i) Model 1 and (ii) Model 2 (100 simulations)

Number of groups	Number of landmarks	Model 1			
		Cumulative variance		Error of reconstructed data	
		GM	FDGM	GM	FDGM
3	20	0.379 (0.022)	0.994 (0.004)	1.229 (0.012)	1.423 (0.025)
	50	0.350 (0.017)	0.997 (0.021)	1.289 (0.0014)	1.411 (0.015)
	100	0.337 (0.024)	0.998 (0.001)	1.314 (0.0112)	1.408 (0.012)
4	20	0.269 (0.017)	0.997 (0.002)	1.816 (0.017)	1.966 (0.019)
	50	0.271 (0.017)	0.996 (0.002)	1.814 (0.019)	1.963 (0.021)
	100	0.270 (0.015)	0.996 (0.002)	1.813 (0.016)	1.963 (0.019)
5	20	0.207 (0.013)	0.996 (0.002)	2.362 (0.018)	2.534 (0.021)
	50	0.208 (0.012)	0.996 (0.002)	2.364 (0.022)	2.537 (0.025)
	100	0.210 (0.012)	0.996 (0.002)	2.367 (0.021)	2.540 (0.023)

(i)

Table 5.1, continued.

Number of groups	Number of landmarks	Model 2			
		Cumulative variance		Error of reconstructed data	
		GM	FDGM	GM	FDGM
3	20	0.947 (0.006)	0.958 (0.0060)	1.222 (0.1173)	0.795 (0.056)
	50	0.948 (0.006)	0.955 (0.008)	1.240 (0.136)	0.815 (0.058)
	100	0.950 (0.006)	0.955 (0.007)	1.238 (0.112)	0.812 (0.060)
4	20	0.947 (0.005)	0.957 (0.005)	1.385 (0.130)	0.810 (0.056)
	50	0.984 (0.005)	0.954 (0.005)	1.391 (0.137)	0.813 (0.051)
	100	0.949 (0.005)	0.953 (0.005)	1.409 (0.127)	0.830 (0.051)
5	20	0.947 (0.005)	0.958 (0.006)	1.372 (0.112)	0.814 (0.058)
	50	0.949 (0.005)	0.955 (0.005)	1.374 (24.746)	7.167 (6.193)
	100	0.949 (0.005)	0.954 (0.006)	1.377 (0.121)	0.825 (0.052)

(ii)

Table 5.2: Mean of classification rate of LDA and FLDA for test data of Model 1 and Model 2 (100 simulations)

Number of groups	Number of landmarks	Model 1		Model 2	
		GM	FDGM	GM	FDGM
3	20	0.376 (0.056)	0.323 (0.059)	0.228 (0.073)	0.333 (0.115)
	50	0.376 (0.056)	0.323 (0.059)	0.180 (0.037)	0.357 (0.093)
	100	0.390 (0.137)	0.352 (0.093)	0.204 (0.059)	0.371 (0.052)
4	20	0.260 (0.042)	0.246 (0.061)	0.128 (0.039)	0.282 (0.037)
	50	0.260 (0.042)	0.278 (0.074)	0.167 (0.044)	0.239 (0.055)
	100	0.232 (0.027)	0.278 (0.110)	0.150 (0.050)	0.246 (0.044)
5	20	0.270 (0.070)	0.250 (0.014)	0.131 (0.030)	0.203 (0.039)
	50	0.300 (0.028)	0.190 (0.042)	0.114 (0.042)	0.202 (0.054)
	100	0.211 (0.034)	0.202 (0.072)	0.105 (0.034)	0.149 (0.034)

Table 5.3: Mean of classification rate of classifiers for (i) GM and (ii) FDGM methods for Model 1 (100 simulations)

Number of groups	Number of landmarks	Model 1				
		GM				
		NB	SVM	RF	GLM	ANN
3	20	0.393 (0.061)	0.358 (0.048)	0.396 (0.060)	0.234 (0.038)	0.368 (0.102)
	50	0.393 (0.061)	0.358 (0.048)	0.396 (0.060)	0.234 (0.038)	0.368 (0.102)
	100	0.463 (0.053)	0.409 (0.068)	0.419 (0.039)	0.244 (0.058)	0.374 (0.086)
4	20	0.328 (0.023)	0.304 (0.051)	0.283 (0.050)	0.173 (0.031)	0.273 (0.052)
	50	0.373 (0.043)	0.330 (0.043)	0.335 (0.047)	0.180 (0.039)	0.242 (0.075)
	100	0.371 (0.053)	0.364 (0.042)	0.347 (0.040)	0.145 (0.036)	0.295 (0.039)
5	20	0.286 (0.028)	0.226 (0.056)	0.220 (0.047)	0.166 (0.047)	0.240 (0.018)
	50	0.300 (0.028)	0.253 (0.000)	0.253 (0.018)	0.193 (0.028)	0.246 (0.009)
	100	0.304 (0.058)	0.266 (0.052)	0.289 (0.055)	0.144 (0.017)	0.222 (0.053)

(i)

Table 5.3, continued.

Number of groups	Number of landmarks	Model 1				
		FDGM				
		NB	SVM	RF	GLM	ANN
3	20	0.374 (0.084)	0.371 (0.080)	0.349 (0.073)	0.250 (0.073)	0.266 (0.044)
	50	0.374 (0.084)	0.371 (0.080)	0.349 (0.073)	0.250 (0.073)	0.266 (0.044)
	100	0.390 (0.042)	0.368 (0.051)	0.333 (0.031)	0.244 (0.028)	0.298 (0.054)
4	20	0.295 (0.051)	0.304 (0.063)	0.245 (0.071)	0.197 (0.060)	0.247 (0.073)
	50	0.269 (0.057)	0.288 (0.052)	0.238 (0.079)	0.161 (0.035)	0.223 (0.046)
	100	0.264 (0.059)	0.250 (0.048)	0.230 (0.059)	0.159 (0.025)	0.185 (0.048)
5	20	0.293 (0.037)	0.260 (0.028)	0.246 (0.065)	0.213 (0.018)	0.206 (0.028)
	50	0.240 (0.056)	0.253 (0.018)	0.213 (0.075)	0.173 (0.018)	0.266 (0.113)
	100	0.264 (0.058)	0.238 (0.026)	0.217 (0.038)	0.150 (0.018)	0.181 (0.035)

(ii)

Table 5.4: Mean of classification rate of classifiers for (i) GM and (ii) FDGM methods for Model 2 (100 simulations)

Number of groups	Number of landmarks	Model 2				
		GM				
		NB	SVM	RF	GLM	ANN
3	20	0.333 (0.115)	0.161 (0.084)	0.038 (0.052)	0.447 (0.131)	0.361 (0.100)
	50	0.285 (0.137)	0.152 (0.092)	0.019 (0.032)	0.371 (0.085)	0.323 (0.071)
	100	0.352 (0.137)	0.171 (0.075)	0.009 (0.025)	0.419 (0.113)	0.295 (0.065)
4	20	0.314 (0.074)	0.114 (0.113)	0.000 (0.000)	0.314 (0.092)	0.323 (0.118)
	50	0.285 (0.084)	0.104 (0.075)	0.010 (0.025)	0.428 (0.126)	0.352 (0.099)
	100	0.323 (0.080)	0.190 (0.080)	0.009 (0.025)	0.428 (0.126)	0.333 (0.121)
5	20	0.266 (0.121)	0.085 (0.074)	0.000 (0.000)	0.419 (0.125)	0.342 (0.151)
	50	0.276 (0.089)	0.076 (0.071)	0.000 (0.000)	0.361 (0.148)	0.304 (0.100)
	100	0.295 (0.093)	0.085 (0.063)	0.000 (0.000)	0.314 (0.083)	0.342 (0.104)

(i)

Table 5.4, continued.

Number of groups	Number of landmarks	Model 2				
		FDGM				
		NB	SVM	RF	GLM	ANN
3	20	0.742 (0.286)	0.752 (0.226)	0.847 (0.161)	0.780 (0.191)	0.695 (0.191)
	50	0.752 (0.125)	0.590 (0.124)	0.704 (0.143)	0.657 (0.146)	0.676 (0.178)
	100	0.666 (0.419)	0.695 (0.297)	0.647 (0.125)	0.600 (0.282)	0.771 (0.217)
4	20	0.733 (0.224)	0.714 (0.074)	0.752 (0.183)	0.800 (0.066)	0.542 (0.135)
	50	0.752 (0.220)	0.714 (0.157)	0.714 (0.236)	0.647 (0.230)	0.647 (0.170)
	100	0.790 (0.190)	0.723 (0.160)	0.819 (0.175)	0.714 (0.147)	0.695 (0.210)
5	20	0.723 (0.141)	0.723 (0.111)	0.828 (0.206)	0.695 (0.230)	0.695 (0.148)
	50	0.780 (0.179)	0.781 (0.226)	0.723 (0.156)	0.695 (0.158)	0.714 (0.183)
	100	0.571 (0.132)	0.542 (0.186)	0.667 (0.172)	0.657 (0.089)	0.666 (0.158)

(ii)

5.4 Application to Real Data

5.4.1 Data Description

The kangaroo landmark dataset is described in detail in Butler et al. (2021). 48 landmarks on crania of 41 extant were used in this study to observe the credibility of the FDGM approach. There are 30 fixed landmarks, placed at "homologous" points on the crania and three sets of semi landmarks, equally spaced along the left molar row (six semi landmarks), right molar row (six semi landmarks), and sagittal axis of the cranial roof (six semi landmarks). To avoid human error in landmarking, the process was repeated twice for each specimen to obtain the mean shape of the two replicates for subsequent analysis (Butler et al., 2021).

Statistical analysis to observe four dietary groups (fungivore, browser, grazer, and mixed feeder) of the extant species was performed in R version 4.2.1. To use the geometric morphometric data, the raw coordinates obtained from the cranial landmarks were aligned using generalised Procrustes analysis (GPA) (Rohlf & Slice, 1990) for optimal registration using translation, rotation, and scaling using the *gpagen* function in the *geomorph* package (Adams & Otárola-Castillo, 2013). Based on Butler et al. (2021), each semi landmark was allowed to slide along its respective tangent directions according to the TPS method (Gunz et al., 2005; Kraatz et al., 2015). The resulting symmetric shape data was used to perform PCA and LDA for both GM and FDGM methods. According to McCane (2013), outline methods produce useful and valid results when suitably constrained by landmarks, which leads to the main idea of this work to incorporate FDA approach (Figure 5.5) to observe the separation among the four dietary groups of the kangaroo extant species.

A total of 16 MFPCs are obtained from the converted functional data using MFPCA. The first three MFPCs explained 77.48% of the total variation in the dietary classification among the marsupials. A distinct cluster among dietary categories when using the MFPC scores (Figure 5.5(b)) is also observed. The PCA from GM yields 40 PCs, where the first three PCs explained 63.85% of variation (Figure 5.5(a)). There is also no overlapping among the groups when the functional approach is applied.

The number of principal components used is based on a threshold value of 90% of variation explained. 11 PCs were selected based on the GM method and 5 MFPCs using FDGM were used in LDA. The results for both approaches reveal a good separation between the four dietary categories, which are class labels. Based on GM, the percentage of separations achieved by the first discriminant function is 89.0%, second is 8.49 % and the third is 2.52 %. The first discriminant function is higher using the FDGM method

compared to GM, which is 97.34%. It is noticeable that the separations between groups are comparable for the two methods (Figure 5.6)).

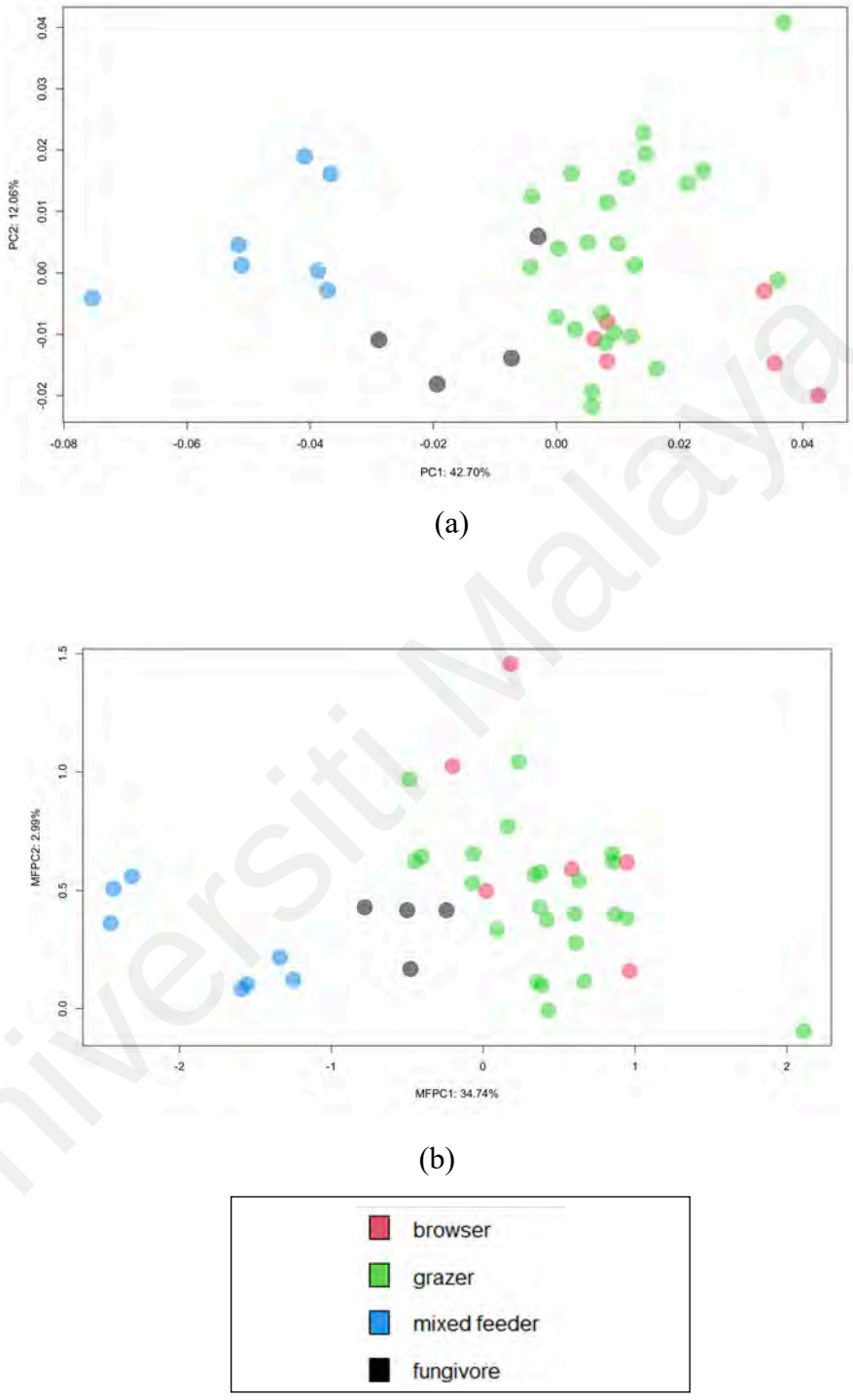


Figure 5.4: The PCs of the (a) GM and (b) FDGM methods for symmetric shape data

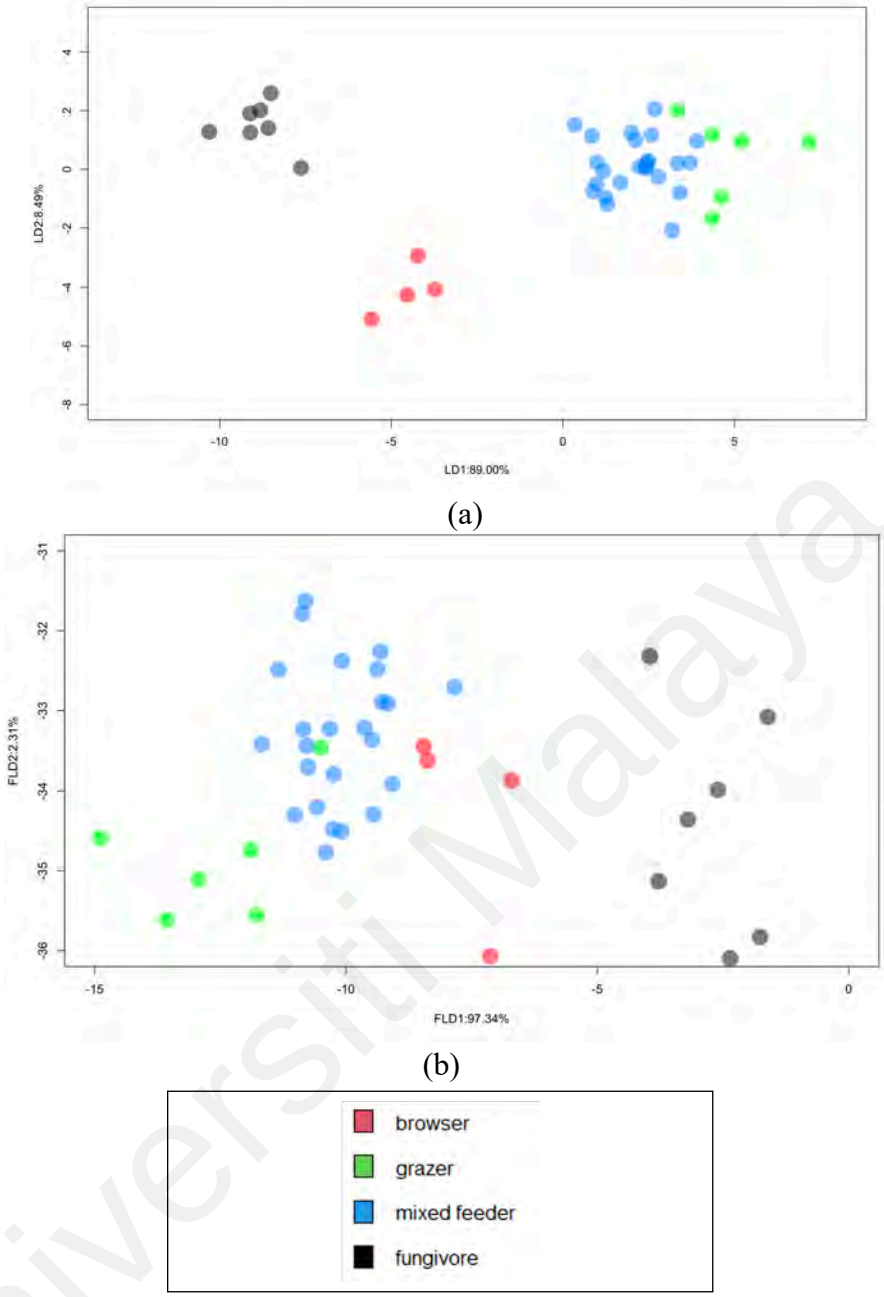


Figure 5.5: The first two LDs of the (a) GM and (b) FDGM methods for symmetric shape data

5.4.2 Results and Discussion

A comparative study of FDGM and GM was done to investigate four dietary categories based on the crania of 41 extant kangaroo species. The findings of this study revealed the existence of the four dietary clusters when the standardised 3D landmarks of the views combined are converted into functional data, rather than being discretised in point sets. The PC scores improve classification as they are projected onto orthonormal eigenfunctions to capture prominent directions. The application of MFPCA reduces dimensionality by projecting the functional landmark data onto the set of orthonormal basis functions which induces the uniqueness of the MFPCA scores for each observation to improve classification accuracy. Using classical landmark-based approach can be difficult to standardise the selections, which can drastically differ results in classification tasks (Srivastava & Klassen, 2016). Thus, FDGM can be a more natural solution as the boundaries of the objects are treated as continuous curves, thus better matches the features across curves (Srivastava & Klassen, 2016).

As shown in Figure 5.5 (a), PCA based on GM gives a comparable presentation to FDGM (Figure 5.5(b)) in terms of the structure variability. The grazing kangaroos group overlaps with the mixed feeding kangaroos' group on all PC axes in both GM and FDGM methods. Based on Figure 5.6, LDA shows a better separation between both methods compared to the PCA results. FLDA forms a linear combination based on class labels to determine the directions of maximum variance, which makes it well-suited for classification tasks. Similar to LDA using the GM method, the FLDA uses linear combinations of continuous functions obtained from the functional data of the 3D landmarks to produce canonical functions to represent the typical LDA setting based on the dietary categories (Gardner-Lubbe, 2021). The MFPCA scores of the selected components are used to perform FLDA.

5.5 Conclusion

This study proposed the FDGM approach on 3D landmark data to represent the shapes of skulls which is an extension of the 2D FDGM framework. Simulation studies and application to real data showed that FDGM performed better than GM when PCA and LDA were employed. FDA methods can be used to analyse shape data as functional curves, which represent the continuous variation in shape across individuals or samples. FDA provides a powerful and flexible framework for analysing shape variation in geometric morphometrics research. This can help researchers to gain new insights into the underlying biological processes and functional relationships between shape and other variables. Outline analysis using FDA approach on the images can be considered for future studies to improve classification accuracy. It is also of interest to overcome the challenge highlighted by White et al. (2023) relating multivariate data response specifications for traits to functional data response specifications to allow relational inference between responses in the search of causal factors in analysing shape.

CHAPTER 6: CONCLUDING REMARKS

6.1 Summary of Findings

In this thesis, RFE was applied into the study of traditional morphometrics to study selecting the most discriminant features for both male and female *Rattus rattus* to reduce the computational complexity of the models for classification of age groups. ANN was provided the best accuracy among three predictive classification models using all features and the RFE-selected features based on the age groups. This study also introduced a novel FDA approach for morphometrics in 2D and 3D geometric morphometrics. FDGM was proposed to classify the three shrew species based on the three craniodental views using 2D landmark data by converting the 2D landmark data into continuous curves, which are then represented as linear combinations of basis functions. Its performance was then compared with GM and the results revealed that FDGM yields comparable results as GM in classifying the three shrew species, and the dorsal view was the best craniodental view for distinguishing the three shrew species. The work also showed that GLM was the most accurate among the five classification models based on the predicted PC scores obtained from both methods (combination of all three craniodental views and individual views). The FDGM method was further extended to the study of 3D landmark data to distinguish dietary categories of kangaroos. Based on the results obtained from the simulation studies conducted for 2D and 3D landmark data and application to real data, this study suggests that FDGM has potential in morphometrics studies to improve the accuracy and resolution of shape variation. Additionally, the use of machine learning algorithms in conjunction with FDGM can further enhance the performance of morphometric studies.

6.2 Contributions

The study has contributed to morphometric studies in the following ways:

i. The application of RFE as an alternative method for selecting features, especially when used alongside PCA, may improve the effectiveness and efficiency of classification models. This combination can leverage the strengths of both methods which is PCA's ability to reduce redundancy and RFE's focus on feature importance, leading to improved model performance in terms of both accuracy and efficiency.

ii. The development of an FDA approach for morphometrics represents enables a more exhaustive and informative analysis of shape variation, making the results of FDGM comparable to those obtained through GM in this study.

iii. FDGM is a new and powerful method for analysing shape variation in 2D and 3D landmark coordinate data. It has the potential to revolutionise the way that morphometric studies are conducted. In addition, this study also highlights the potential of machine learning algorithms for enhancing the performance of morphometric studies.

6.3 Further Research

As decision trees inherently perform feature selection, future studies could benefit from comparing the RFE approach with decision tree methods to evaluate their effectiveness in morphometric analyses. The FDGM can be further refined by addressing issues related to landmark coordinate data analysis, specifically by eliminating nuisance parameters such as translation and rotation (Lele and McCulloch, 2002).

Conducting intensive simulations on both regular and irregular data, with the introduction of hyperparameter selection methods and with different datasets among several populations, species, and individuals, would be an intriguing area of study. This approach could be extended to image analysis by directly considering the specimen outlines from images and applied to a broader range of biological organisms, including plants, animals, and microorganisms. Moreover, the FDGM method could be integrated with various biological research disciplines, such as evolutionary biology, developmental biology, and ecology, to gain new insights into the biological significance of shape variation. It could also be extended to artificial intelligence (AI) for automatic recognition of organisms, incorporating categorical functional data. For instance, FDGM could be used to investigate how shape variation relates to environmental adaptation or is influenced by different developmental processes. Additionally, future morphometric research could benefit from developing and applying new machine learning algorithms based on the FDGM method, potentially leading to significant advancements in the field.

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.

Abu, A., Leow, L. K., Ramli, R., & Omar, H. (2016). Classification of *Suncus murinus* species complex (*Soricidae: Crocidurinae*) in Peninsular Malaysia using image analysis and machine learning approaches. *BMC Bioinformatics*, *17*(19), 107-118.

Adams, D. C., & Otárola-Castillo, E. (2013). geomorph: an r package for the collection and analysis of geometric morphometric shape data. *Methods in Ecology and Evolution*, *4*(4), 393–399.

Adams, D. C., Collyer, M. L., & Kaliontzopoulou, A. (2018). Geomorph: Software for geometric morphometric analyses. R package version 3.0.6.

Adams, D. C., Rohlf, F. J., & Slice, D. E. (2004). Geometric morphometrics: Ten years of progress following the 'revolution.' *Italian Journal of Zoology*, *71*(1), 5–16.

Adams, D., Rohlf, F., & Slice, D. (2013). A field comes of age: Geometric morphometrics in the 21st century. *Hystrix, the Italian Journal of Mammalogy, 24*(1), 7-14.

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.

Albrecht, G. H. (1979). The study of biological versus statistical variation in multivariate morphometrics: the descriptive use of multiple regression analysis. *Systematic Biology*, *28*(3), 338–344.

Arai, Y., Kanaiwa, M., Kato, M., & Kobayashi, M. (2021). Morphological identification in skull between spotted seal and harbor seal using geometric morphometrics. *Journal of Morphology*, *282*(10), 1455–1465.

Arias-Martorell, J., Alba, D., Potau, J. M., Bello-Hellegouarch, G., & Pérez-Pérez, A. (2015). Morphological affinities of the proximal humerus of *Epipliopithecus vindobonensis* and *Pliopithecus antiquus*: Suspensory inferences based on a 3D geometric morphometrics approach. *Journal of Human Evolution*, *80*, 83–95.

Astúa de Moraes, D., Hingst-Zaher E., Marcus, L., Cerqueira, R. (2000). A geometric morphometric analysis of cranial and mandibular shape variation of *didelphid* marsupials. *Hystrix, the Italian Journal of Mammalogy*, *11*(1),1-4140.

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.

Beck, R. M. D., Voss, R. S., & Jansa, S. A. (2022). Craniodental morphology and phylogeny of marsupials. *Bulletin of the American Museum of Natural History*, *457*,1-352.

Bellin, N., Calzolari, M., Callegari, E., Bonilauri, P., Grisendi, A., Dottori, M., & Rossi, V. (2021). Geometric morphometrics and machine learning as tools for the identification of sibling mosquito species of the *Maculipennis* complex (*Anopheles*). *Infection, Genetics and Evolution*, *95,* 105034.

Berio, F., Bayle, Y., Baum, D., Goudemand, N., & Debiais-Thibaud, M. (2022). Hide and seek shark teeth in random forests: machine learning applied to *Scyliorhinus canicula* populations. *PeerJ*, *10, e*13575.

Bermejo, J. F., Fernández, J. F. G., Polo, F. O., & Márquez, A. C. (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, 9*(9), 1844.

Birkby, W. H. (1966). An evaluation of race and sex identification from cranial measurements. *American Journal of Physical Anthropology*, *24*(1), 21–27.

Blackith R., & Reyment R. A. (1971). *Multivariate morphometrics*. New York, NY: Academic Press.

Bookstein, F. L. (1984). A statistical method for biological shape comparisons. *Journal of Theoretical Biology*, *107*(3), 475–520.

Bookstein, F. L. (1986). Size and shape spaces for landmark data in two dimensions. *Statistical Science*, *1*(2)*,* 181-222.

Bookstein, F. L. (1987). Describing a craniofacial anomaly: Finite elements and the biometrics of landmark locations. *American Journal of Physical Anthropology*, *74*(4), 495–509.

Bookstein, F. L. (1991). Morphometric tools for landmark data. Cambridge: Cambridge University Press.

Bookstein, F. L. (1996). Biometrics, biomathematics and the morphometric synthesis. *Bulletin of Mathematical Biology*, *58*(2), 313–365.

Bookstein, F. L. (1997). Landmark methods for forms without landmarks: morphometrics of group differences in outline shape. In *Medical Image Analysis*, *1*(3), 225-243.

Bookstein, F. L. (1998). A hundred years of morphometrics. *Acta Zoologica Academiae Scientiarum Hungaricae*, *44*(1-2), 7–59.

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.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2), 123-140.

Breiman, L. (2001). Random forests. *Machine Learning*. 45, 5–32.

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.

Butler, K., Travouillon, K. J., Evans, A. R., Murphy, L., Price, G. J., Archer, M., … Weisbecker, V. (2021). 3D morphometric analysis reveals similar ecomorphs for early kangaroos (*macropodidae*) and fanged kangaroos (*balbaridae*) from the Riversleigh world heritage area, Australia. *Journal of Mammalian Evolution*, *28*(2), 199–219.

Caple, J., Byrd, J., & Stephan, C. N. (2017). Elliptical fourier analysis: fundamentals, applications, and value for forensic anthropology. *International Journal of Legal Medicine*, *131*(6), 1675–1690.

Cardini, A., & Elton, S. (2008). Does the skull carry a phylogenetic signal? Evolution and modularity in the guenons. *Biological Journal of the Linnean Society*, *93*(4), 813–834.

Chen, K., Chen, K., Müller, H.-G., & Wang, J.L. (2011). Stringing high-dimensional data for functional analysis. *Journal of the American Statistical Association*, *106*(493), 275–284.

Chiaverini, L., Macdonald, D. W., Hearn, A. J., Kaszta, Ż., Ash, E., Bothwell, H. M., … Cushman, S. A. (2023). Not seeing the forest for the trees: Generalised linear model out-performs random forest in species distribution modelling for Southeast Asian felids. *Ecological Informatics*, *75*, 102026.

Christopher, M., & John, R. (2022). *Package "MLeval" Title Machine Learning Model Evaluation*.

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.

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.

de Almeida, V. E., de Sousa Fernandes, D. D., Diniz, P. H. G. D., de Araújo Gomes, A., Véras, G., Galvão, R. K. H., & Araujo, M. C. U. (2021). Scores selection via Fisher's discriminant power in PCA-LDA to improve the classification of food data. *Food Chemistry*, *363*, 130296.

Denisko, D., & Hoffman, M. M. (2018). Classification and interaction in random forests. *Proceedings of the National Academy of Sciences*, *115*(8), 1690–1692.

Dryden, I. L., & Mardia, K. V. (2016). *Statistical Shape Analysis, with Applications in R*. Wiley.

Dryden, I.L., & Mardia, K. V. (1998). *Statistical shape analysis*. John Wiley and Sons, Chichester.

Dudzik, B. (2019). Examining cranial morphology of Asian and Hispanic Populations using geometric morphometrics for ancestry estimation. *Forensic Anthropology*, *2*(4).

Dujardin, J. P. (2017). Modern morphometrics of medically important arthropods. In *Genetics and Evolution of Infectious Diseases: Second Edition* (pp. 285–311). New York, NY: Elsevier Inc.

Dujardin, J.P., Kaba, D., Solano, P., Dupraz, M., McCoy, K. D., & Jaramillo-O, N. (2014). Outline-based morphometrics, an overlooked method in arthropod studies? *Infection, Genetics and Evolution*, *28*, 704–714.

Epifanio, I., & Ventura-Campos, N. (2011). Functional data analysis in shape analysis. *Computational Statistics & Data Analysis*, *55*(9), 2758–2773.

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.

Ferraty, F., & Vieu, P. (2006). *Nonparametric functional data analysis [Theory and practice]*. New York: Springer.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*(2), 179–188.

Francis, C. (2008). A guide to the mammals of south-east Asia. In *Journal of Mammalogy*. London, England: New Holland Publishers.

Gardner-Lubbe, S. (2021). Linear discriminant analysis for multiple functional data analysis. *Journal of Applied Statistics*, *48*(11), 1917–1933.

Gasser, T., & Kneip, A. (1995). Searching for structure in curve sample. *Journal of the American Statistical Association*, *90*(432), 1179.

Giles, E., & Elliot, O. (1963). Sex determination by discriminant function analysis of crania. *American Journal of Physical Anthropology*, *21*(1), 53–68.

Gower, J. C. (1975). Generalized procrustes analysis. *Psychometrika*, *40*(1), 33–51.

Gunz, P., Mitteroecker, P., & Bookstein, F. L. (2005). Semilandmarks in three dimensions. In *Modern Morphometrics in Physical Anthropology* (pp. 73–98). New York, NY: Kluwer Academic Publishers-Plenum Publishers.

Guo, X., Wu, W., & Srivastava, A. (2022). Data-Driven, Soft Alignment of Functional Data Using Shapes and Landmarks. *ArXiv*.

Hall, P., & Vial, C. (2006). Assessing the finite dimensionality of functional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *68*(4), 689–705.

Happ, C., & Greven, S. (2018). multivariate functional principal component analysis for data observed on different (dimensional) domains. *Journal of the American Statistical Association*, *113*(522), 649–659.

Happ-Kurz, C. (2020). Object-oriented software for functional data. *Journal of Statistical Software*, *93*(5).

Horgan, G. W. (2000). Principal component analysis of random particles. *Journal of Mathematical Imaging and Vision, 12*, 169175.

Horváth, L., & Kokoszka, P. (2012). *Inference for Functional Data with Applications* (pp. 37–39). New York, NY: Springer.

Howells, W.W. (1989). Skull shapes and the map. craniometric analyses in the dispersion of modern homo. Cambridge, Peabody Museum of Archaeology and Ethnology, Harvard University.

Hutterer, R. (2005). Order *Soricomorpha*. In *Mammal species of the world: A taxonomic and geographic reference, 88*(3), 824-830.

IBM Corp. Released 2020. IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp.

Jamaluddin, S. A., Rahman, M. R. A., Othman, N., Haris, H., Zahari, F. N. A., Najmuddin, M. F., … Abdul-Latiff, M. A. B. (2022). Diversity of non-volant small mammals in Pulau Tinggi, Johor, Malaysia. *Journal of Sustainability Science and Management*, *17*(11), 121–129.

James, G. M., & Hastie, T. J. (2001). Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *63*(3), 533–550.

James Rohlf, F., & Marcus, L. F. (1993). A revolution morphometrics. *Trends in Ecology & Evolution*, *8*(4), 129–132.

Kassambara, A., & Mundt, F. (2020). *Extract and Visualize the Results of Multivariate Data Analyses [R package factoextra version 1.0.7]*.

Kendall, D. G. (1984). Shape manifolds, Procrustean metrics, and complex projective spaces. *Bulletin of the London Mathematical Society*, *16*(2), 81–121.

Khang, T. F., Mohd Puaad, N. A. D., Teh, S. H., & Mohamed, Z. (2021). Random forests for predicting species identity of forensically important blow flies (*Diptera: Calliphoridae*) and flesh flies (*Diptera: Sarcophagidae*) using geometric morphometric data: Proof of concept. *Journal of Forensic Sciences*, *66*(3), 960–970.

Klingenberg, C. P. (2011). MorphoJ: an integrated software package for geometric morphometrics. *Molecular Ecology Resources, 11,* 353-357.

Kneip, A., & Gasser, T. (1992). Statistical tools to analyze data representing a sample of curves. *The Annals of Statistics*, *20*(3), 1266–1305.

Kononenko, I. (1990). *Comparison of inductive and naive Bayesian learning approaches to automatic knowledge acquisition.* Amsterdam: IOS Press.

Kraatz, B. P., Sherratt, E., Bumacod, N., & Wedel, M. J. (2015). Ecological correlates to cranial morphology in Leporids (Mammalia, Lagomorpha). *PeerJ*, *3*, e844.

Kuhl, F. P., & Giardina, C. R. (1982). Elliptic fourier features of a closed contour. In *Computer Graphics and Image, 18*(3), 236-258.

Kuhn, M. (2008). Building predictive models in *R* using the caret package. *Journal of Statistical Software*, *28*(5),1-26.

Langley, P., Iba, W., & Thompson, K. (1992). An analysis of Bayesian classifiers. *Proceedings of the Tenth National Conference on Artificial intelligence, 90*, 223-228.

Marcus L.F. (1990). Traditional morphometrics. *In Proceedings of the Michigan Morphometrics Workshop* (pp.77–122). USA: The University of Michigan Museum of Zoology Ann Arbor.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2016). Feature selection: A data perspective. *ACM Computing Surveys*, *50*(6), 1–45.

Libois, R., Ramalhinho, G., Mathias, M. da L., Santos-Reis, M., Fons, R., Petrucci-Fonseca, F., … M., C.-P. (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*, 245–251.

MacLeod, N. (2007). *Automated Taxon Identification in Systematics: Theory, Approaches and Applications*. *Systematics Association Special Volumes*. Boca Raton, FL: CRC Press.

MacLeod, N. (2013). Landmarks and semilandmarks: Differences without meaning and meaning without difference. *Palaeontological Association Newsletter* ,*82*, 32–43.

Macleod, N. (2017). On the use of machine learning in morphometric analysis. *Biological Shape Analysis*, 134–171. World Scientific. PE Lestrel, editor. Biological shape analysis: proceedings of the 4th international symposium. Hackensack, NJ: World Scientific. p. 134–71.

Maderbacher, M., Bauer, C., Herler, J., Postl, L., Makasa, L., & Sturmbauer, C. (2008). Assessment of traditional versus geometric morphometrics for discriminating populations of the *Tropheus moorii* species complex (*Teleostei: Cichlidae*), a Lake Tanganyika model for allopatric speciation. *Journal of Zoological Systematics and Evolutionary Research*, *46*(2), 153–161.

Marcy, A. E., Fruciano, C., Phillips, M. J., Mardon, K., & Weisbecker, V. (2018). Low resolution scans can provide a sufficiently accurate, cost- and time-effective alternative to high resolution scans for 3D shape analyses. *PeerJ*, *6*, e5032.

Martensson, T. (1998). Measurement error in geometric morphometrics: empirical strategies to assess and reduce its impact on measures of shape. *In Acta Zoologica Academiae Scientiarum Hungaricae* ,44(1-2),1-6.

Martinez, A. M., & Kak, A. C. (2001). PCA versus LDA. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(2), 228–233.

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*,29, 617–663.

McCane, B. (2013). Shape variation in outline shapes. *Systematic Biology*, *62*(1), 134–146.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4), 115-133.

Misra, P., & Singh Yadav, A. (2020). Improving the classification accuracy using recursive feature elimination with cross-validation. *International Journal on Emerging Technologies*, *11*(3), 659–665.

Mitteroecker, P., & Gunz, P. (2009). Advances in geometric morphometrics. *Evolutionary Biology*, *36*(2), 235–247.

Mitteroecker, P., & Schaefer, K. (2022). Thirty years of geometric morphometrics: Achievements, challenges, and the ongoing quest for biological meaningfulness. *American Journal of Biological Anthropology*, *178*(S74), 181–210.

Zelditch M.L., Swiderski, D.L., Sheets, H. D., & Fink W.L., (2004). *Geometric Morphometrics of Biologists: A Primer*. *Elsevier Academic Press*. New York and London.

Mohamad Ikbal, N. H., Pathmanathan, D., Bhassu, S., Simarani, K., & Omar, H. (2019). Morphometric analysis of craniodental characters of the house rat, *Rattus rattus* (*Rodentia: Muridae*) in Peninsular Malaysia. *Sains Malaysiana*, *48*, 2103–2111.

Morton, J. T., Toran, L., Edlund, A., Metcalf, J. L., Lauber, C., & Knight, R. (2017). Uncovering the horseshoe effect in microbial analyses. *MSystems*, *2*(1).

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*). In *The Raffles Bulletin of Zoology* ,*52,* 653-663.

Musser, G. G., & Newcomb, Cameron. (1983). *Malaysian murids and the giant rat of Sumatra. Bulletin of the AMNH ; 174(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*, 1–41.

Mustaqeem, M., & Saqib, M. (2021). Principal component based support vector machine (PC-SVM): a hybrid technique for software defect detection. *Cluster Computing*, *24*(3), 2581–2595.

Nichols, J. A., Herbert Chan, H. W., & Baker, M. A. B. (2019, February 7). Machine learning: applications of artificial intelligence to imaging and diagnosis. *Biophysical Reviews*, Vol. 11, pp. 111–118. Springer Verlag.

Omar, H., Hashim, R., Bhassu, S., & Ruedi, M. (2013). Morphological and genetic relationships of the *Crocidura monticola* species complex (*Soricidae: Crocidurinae*) in Sundaland. *Mammalian Biology*, *78*(6), 446–454.

Ousley, S. D. (2004). Programme 3Skull.

Podani, J., & Miklos, I. (2002). Resemblance coefficients and the horseshoe effect in principal coordinates analysis. *Ecology*, *83*(12), 3331.

Prajwala T R. (2015). A comparative study on decision tree and random forest using R tool. *IJARCCE*, 196–199.

Ramsay, J. O. (2006). Functional data analysis. In *Encyclopedia of Statistical Sciences*. Hoboken, NJ, USA: John Wiley & Sons, Inc.

Ramsay, J. O., & Li, X. (1998). Curve registration. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *60*(2), 351–363.

Ramsay, J. O., & Silverman, B. W. (2005). *Functional Data Analysis*. New York, NY: Springer New York.

Rasmussen, C. E. (2004). *Gaussian Processes in Machine Learning*. The MIT Press.

Robinson, D. L., Blackwell, P. G., Stillman, E. C., & Brook, A. H. (2002). Impact of landmark reliability on the planar Procrustes analysis of tooth shape. In *Archives of Oral Biology*, *47 (7)*, 545-554.

Rodrigues, P. J., Gomes, W., & Pinto, M. A. (2022). DeepWings: automatic wing geometric morphometrics classification of honey bee (*Apis mellifera*) subspecies using deep learning for detecting landmarks. *Big Data and Cognitive Computing*, *6*(3),70.

Rohlf, F.J. (1990). Morphometrics. *Annual Review of Ecology and Systematics*, *21*, 299–316.

Rohlf, F. J. (2017). *TpsDig (Version 2.31)*.

Rohlf, F J. (1972). Blackith R. E. and Reyment R. A., Multivariate morphometrics. *Systematic Biology*, *21*(3), 348–349.

Rohlf, F. J. (2000). Statistical power comparisons among alternative morphometric methods. *American Journal of Physical Anthropology*, *111*(4), 463–478.

Rohlf, F. J., & Slice, D. (1990). Extensions of the Procrustes method for the optimal superimposition of landmarks. *Systematic Zoology*, *39*(1), 40.

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statisical Computing. Vienna, Austria.

Rojas, R. (1996). Neural Networks: A Systematic Introduction. Berlin: Springer-Verlag.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386-408.

Ruedi, M., Courvoisier, C., Vogel, P., & Catzeflis, F. M. (1996). Genetic differentiation and zoogeography of Asian *Suncus murinus* (*Mammalia: Soricidae*). In *Biological Journal of the Linnean Society*,*57(4)*, 307–316.

Rummel, A. D., Sheehy, E. T., Schachner, E. R., & Hedrick, B. P. (2024). Sample size and geometric morphometrics methodology impact the evaluation of morphological variation. *Integrative Organismal Biology*, *6*(1), obae002.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533-536.

Salifu, D., Ibrahim, E. A., & Tonnang, H. E. Z. (2022). Leveraging machine learning tools and algorithms for analysis of fruit fly morphometrics. *Scientific Reports*, *12*(1), 7208.

Sammut, C, & Webb, G. (2010). *Encyclopedia of Machine Learning* (Claude Sammut & G. I. Webb, Eds.). Boston, MA: Springer US.

Schunke, A. C., Bromiley, P. A., Tautz, D., & Thacker, N. A. (2012). TINA manual landmarking tool: software for the precise digitization of 3D landmarks. Frontiers in zoology, 9(1), 1-6.

Sheets, H. D., & Webster, M. (2010). A practical introduction to landmark-based geometric morphometrics. *The Paleontological Society Papers*, *16*, 163–188.

Slice, D. E. (2005). Modern morphometrics. In D. E. Slice (Ed.), *Modern Morphometrics in Physical Anthropology* (pp. 1–45). Boston, MA: Springer US.

Srivastava, A., & Klassen, E. P. (2016). *Functional and Shape Data Analysis*. New York, NY: Springer.

Srivastava, A., Wu, W., Kurtek, S., Klassen, E., & Marron, J. (2011). *Registration of Functional Data Using Fisher-Rao Metric*. *ArXiv*.

Stafford, B. J., & Szalay, F. S. (2000). Craniodental functional morphology and taxonomy of *Dermopterans. Journal of Mammalogy*, *81*(2), 360–385.

Tan, J.W., Chang, S.W., Abdul Kareem, S., Yap, H. J., & Yong, K.T. (2018). Deep learning for plant species classification using leaf vein morphometric. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *17*(1), 82-90.

Tang, Y., Horikoshi, M., & Li, W. (2016). ggfortify: Unified interface to visualize statistical results of popular R packages. *The R Journal*, *8*(2), 474.

Thomas, O. O., Shen, H., Raaum, R. L., Harcourt-Smith, W. E. H., Polk, J. D., & Hasegawa-Johnson, M. (2023). Automated morphological phenotyping using learned shape descriptors and functional maps: A novel approach to geometric morphometrics. *PLoS Computational Biology*, *19*(1), e1009061.

Thomson, J. A. (1917). On growth and form. *Nature*, *100* (2498), 21–22.

Thuiller, W., Araújo, M. B., & Lavorel, S. (2003). Generalized models vs. classification tree analysis: Predicting spatial distributions of plant species at different scales. *Journal of Vegetation Science*, *14*(5), 669–680.

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.

Tse, Y. T., & Calede, J. J. M. (2021). Quantifying the link between craniodental morphology and diet in the *Soricidae* using geometric morphometrics. *Biological Journal of the Linnean Society*, *133*(1), 28–46.

Ullah, S., & Finch, C. F. (2013). *Applications of functional data analysis: A systematic review, 13*(43), 1-12.

van Bemmelen van der Plaat, A., van Treuren, R., & van Hintum, T. J. L. (2021). Reliable genomic strategies for species classification of plant genetic resources. *BMC Bioinformatics*, *22*(1), 173.

Van Valen, L. (1974). Multivariate structural statistic in natural history. *Journal of Theoretical Biology*, *45*(1), 235–247.

Vasil'ev, A. G., Vasil'eva, I. A., & Kourova, T. P. (2015). Analysis of coupled geographic variation of three shrew species from Southern and Northern Ural taxocenes. *Russian Journal of Ecology*, *46*(6), 552–558.

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S*. New York, NY: Springer New York.

Viacava, P., Baker, A. M., Blomberg, S. P., Phillips, M. J., & Weisbecker, V. (2022). Using 3D geometric morphometrics to aid taxonomic and ecological understanding of a recent speciation event within a small Australian marsupial (Antechinus: Dasyuridae). *Zoological Journal of the Linnean Society*, *196*(3), 963–978.

Vilchis-Conde, J. M., Ospina-Garcés, S. M., Ureta, C., Cervantes, F. A., & Guevara, L. (2023). Geometric morphometrics clarifies the taxonomic status of semifossorial shrews (*Eulipotyphla, Soricidae, Cryptotis*) from Mexican cloud forests. *Mammalia*, *87*(5), 518–526.

Wang, J.L., Chiou, J.M., & Mueller, H.G. (2016). Review of functional data analysis. *Annual Review of Statistics and Its Application*, *3*, 257-295.

Watanabe, A. (2018). How many landmarks are enough to characterize shape and size variation? *Plos One*, *13*(6), e0198341.

Webb, G. I. (2011). Naïve Bayes. In Claude Sammut & G. I. Webb (Eds.), *Encyclopedia of Machine Learning* (pp. 713–714). Boston, MA: Springer US.

Webster, M., & Sheets, H. D. (2010). A practical introduction to landmark-based geometric morphometrics. *The Paleontological Society Papers*, *16*, 163–188.

White, P. A., Christensen, M. F., Frye, H., Gelfand, A. E., & Silander, J. A. (2023). Joint multivariate and functional modeling for plant traits and reflectances. *Environmental and Ecological Statistics*, *30*, 501–528.

Wolfer, A., Ebbels, T., & Cheng, J. (2022). *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.

Zhang, M., Ruan, Y., Bai, M., Chen, X., Li, L., Yang, X., … Du, X. (2023). Geometric morphometric analysis of genus *Chaetocnema* (*Coleoptera: Chrysomelidae: Alticini*) with insights on its subgenera classification and morphological diversity. *Diversity*, *15*(8), 918.

Zipunnikov, V., Caffo, B., Yousem, D. M., Davatzikos, C., Schwartz, B. S., & Crainiceanu, C. (2011). Functional principal component model for high-dimensional brain imaging. *NeuroImage*, *58*(3), 772–784.