

1 MORPHOMETRIC AND MORPHOLOGICAL-BASED
2 NON-INVASIVE SEX IDENTIFICATION OF BLOOD
3 COCKLE, *Tegillarca granosa* (LINNAEUS, 1758)

4 A Special Problem

5 Presented to

6 the Faculty of the Division of Physical Sciences and Mathematics

7 College of Arts and Sciences

8 University of the Philippines Visayas

9 Miag-ao, Iloilo

10 In Partial Fulfillment

11 of the Requirements for the Degree of

12 Bachelor of Science in Computer Science by

13 ADRICULA, Briana Jade

14 PAJARILLA, Gliezel Ann

15 VITO, Ma. Christina Kane

16 Francis DIMZON, Ph.D.

17 Adviser

18 Victor Marco Emmanuel FERRIOLS, Ph.D.

19 Co-Adviser

20 May 2025

21

Approval Sheet

22

The Division of Physical Sciences and Mathematics, College of Arts and
Sciences, University of the Philippines Visayas

24

certifies that this is the approved version of the following special problem:

25

**MORPHOMETRIC AND MORPHOLOGICAL-BASED
NON-INVASIVE SEX IDENTIFICATION OF BLOOD
COCKLE, *Tegillarca granosa* (LINNAEUS, 1758)**

26

27

28

Approved by:

29

Name	Signature	Date
Francis D. Dimzon, Ph.D. (Adviser)	_____	_____
Victor Marco Emmanuel N. Ferriols, Ph.D. (Co-Adviser)	_____	_____
Ara Abigail E. Ambita (Panel Member)	_____	_____
Kent Christian A. Castor (Division Chair)	_____	_____

30 Division of Physical Sciences and Mathematics
31 College of Arts and Sciences
32 University of the Philippines Visayas

33 **Declaration**

34 We, Briana Jade Adricula, Gliezel Ann Pajarilla, and Ma. Christina Kane
35 Vito, hereby certify that this Special Problem has been written by us and is the
36 record of work carried out by us. Any significant borrowings have been properly
37 acknowledged and referred.

Name	Signature	Date
Briana Jade D. Adricula (Student)	_____	_____
Gliezel Ann T. Pajarilla (Student)	_____	_____
Ma. Christina Kane B. Vito (Student)	_____	_____

Dedication

40

To our family, advisers, and the people of science:

41

A heart full of love,

42

To those who gave wings so we can fly.

43

Stood firm even through moments of doubt.

44

A jovial harmony and warmth that kept us steadfast.

45

A word of thanks is an understatement,

46

To those who cast their light upon our way.

47

A source of wisdom even when the road grew heavy,

48

A north star that guided us through this journey.

49

Immeasurable esteem we offer,

50

To the unsung heroes of science and innovation,

51

Whose drive and dedication uplift and inspire,

52

Changing lives with boundless determination.

53

Acknowledgment

54 **I. General Acknowledgment**

55

56 The researchers would like to extend their heartfelt gratitude to the significant
57 people who were part of this journey. These people extended their expertise,
58 support, and guidance, which made this whole process bearable and fulfilling.

59 First and foremost, to the Almighty God who gives us the strength to never
60 lose hope throughout the challenging moments in our journey; for the wisdom
61 and perseverance He bestowed upon us to accomplish this paper. All for His
62 unconditional love, mercy, and grace. All glory and praise to Him.

63 To our research adviser who provided guidance and support to the researchers,
64 and for the help when things were hard. Dr. Francis Dimzon – thank you for
65 guiding us throughout this journey. We would like to thank you for assuring
66 us, especially when we were about to turn around halfway. We would like to
67 extend our greatest gratitude for the consistent supervision by providing us with
68 valuable feedback and insightful comments that aided us in our methodology and
69 in completing this paper.

70 To our co-adviser who was willing to tour us around the hatchery facility and
71 helped us come up with this study. Thank you, Dr. Victor Marco Emmanuel
72 Ferriols, for sharing your knowledge and wisdom with us researchers. Thank you
73 for answering our questions and extending your time with us. Additionally, the
74 most important part of the research is the samples — the blood cockle samples
75 — thank you for providing them willingly. You served as an inspiration for us to
76 apply solutions to real-world problems, especially in aquaculture.

77 To the research associates and hatchery staff of the UPV Hatchery, Ms. Allena
78 Arteta, Ms. LC May Gasit, Ms. Shiela Untalan, and the hatchery staff, Mr. Paul
79 Andre Lopez and Mr. Joel M. Fabrigas, for assisting us in handling the blood
80 cockles, teaching us how to identify their sex, spawning, and dissection. A big
81 thanks to Ms. Allena Arteta for letting us borrow the camera stand that we used
82 in the entire data gathering process. Your constant warmth and assistance went
83 beyond learning the basics; instead, we learned a whole lot more, especially the
84 importance of creating solutions for aquaculture practices.

85 Lastly, to everyone who directly or indirectly contributed to this thesis, whether
86 through words of encouragement, technical advice, or simply by believing in us,
87 we offer our sincerest gratitude. This study would also not have been possible
88 without you.

89 **i. Adricula's Acknowledgment**

90
91 I would like to express my deepest gratitude to my beloved family who sup-
92 ported me throughout my academic journey. My Mama Madelene and Papa Juvy,
93 who have made countless sacrifices to make me the person I am today. To my
94 sister, Aianna, and cousins Jayson, Joana, and Joel, who have helped me unwind
95 after long and stressful days of academic load. To my aunt, Tita Lybel – thank
96 you for believing in me since day one. You are always there to guide me in every
97 decision I make. And to my Lola Lydia, who has been my inspiration in finish-
98 ing my studies. Your reminders to eat more and stay safe have always kept me
99 grounded and cared for, and I appreciate the extra allowance you always give me
100 whenever I go back to Miagao. Without such a supportive family behind me, I

101 couldn't have done this without you.

102 To my pretty college friends, Arianne, Gliezel, Karielle, Kane, Kzlyr, and
103 Sharah, thank you for making my college experience bearable. The laughter and
104 joy that we shared together lessened the stress I felt. Here's to all the rants,
105 tsismis, and bangs memories we had — look how far we've come. I will forever
106 cherish the memories we made together through ups and downs. Thank you for
107 the encouragement along the way and just for being there by my side.

108 To my co-researchers, Gliezel and Kane, this would not have been possible
109 without you. Your insights and passion for this research inspired me along the
110 way. Thank you for the teamwork, initiatives, and hard work. I will miss the
111 dorm to hatchery to computer lab to dorm adventures with you!

112 To my second home, UPV Balay Gumamela, our dorm manager, and all the
113 staff — thank you for checking up on me, always making me feel welcome, and
114 ensuring that we had a comfortable stay.

115 And lastly, to the people who believed in me and have always been there for
116 me every step of the way, even when life was harsh — I am forever grateful for
117 your unwavering encouragement, support, and just for believing that I can make
118 things possible.

119 **ii. Pajarilla's Acknowledgment**

120
121 To my mother in heaven, Mama Eva, I made it through the last chapter of
122 this paper. I felt your boundless love and support every step of the way. You will
123 always be my source of strength and inspiration, even when hope seemed bleak

124 and weary. The different storms that were pushed my way — I stood because I
125 know you believed in me and wanted the best for me and my three siblings.

126 Greatest love and appreciation to my siblings – my source of boundless joy.
127 Your smiles, laughter, and warmth gave me a boost of energy that adenosine
128 triphosphate can't surpass. Gil Ivan, Gillian Gella, and my small baby brother
129 Gian Godric – no words can express how thankful I am for your existence and
130 love, as it sparked perseverance and grit in my heart. To the people who cared
131 and looked after me, thank you, Tito Rem and Auntie Maricel, for the financial
132 support and encouragement. Thank you for believing in me and the things that
133 I can achieve.

134 My heart is full of gratitude and appreciation to the Gumamela Girls Dormers.
135 I cherish all the laughter, joy, sadness, ennui, anger, fear, embarrassment, anxiety,
136 envy, and disgust. Kidding aside, I am forever thankful for the bond and friendship
137 we made, the spontaneous moments, and baybay sessions, admiring the beauty of
138 nature. Arianne, Briana, Kane, Karielle, Kzlyr, Sharah – thank you for the words
139 of encouragement, for listening, and for your companionship.

140 I would like to extend my greatest appreciation to my co-researchers, Kane and
141 Briana, who trailed the 100 steps from and to the hatchery facility. I appreciate
142 the moments and laughter that we spent together, from ideation, spending almost
143 the whole day in the computer laboratory to train our model, to crafting this
144 final paper. We went through doubts and a series of burnout moments, but your
145 presence made everything lighter.

146 My love and appreciation are offered to my high school and Senior High School
147 friends who listened to my “I can't do this anymore” moments. Thank you,

148 Jara, Francine, Nellen, Mary, and Nyle, for cheering me on and believing in my
149 potential, even when things sometimes got so dim. Thank you for reminding me
150 that there is always light at the end of the tunnel. You all are my support system,
151 cheering squad, and parents I met along the way.

152 My greatest thanks are relayed to Manang Myrtle, who offered kind words,
153 warmth, and guidance. Thank you so much for the support, for offering your
154 shoulders when I needed someone to rely on. Things felt a bit lighter with your
155 smile and your “You can do it, Gliez.” Thank you so much for being my guardian
156 for a year.

157 My gratitude extends to the people inside the university who served as family.
158 To Ma’am Marites Geonanga, Ma’am Paula Khryss Ushiyama, Balay Gumamela
159 dorm manager, Ma’am Celina Sumalapao and the dorm staff, Tita Ope, Tita
160 Vilma, Nong Joefil, and the UPV Virgils — thank you for the support, for believ-
161 ing in me, and for the nutritional food that helped me brainstorm, conduct, and
162 write this study.

163 To everyone who supported me, thank you so much. Words can’t express how
164 grateful I am.

165 **iii. Vito’s Acknowledgment**

166
167 To my beloved family – Mama, Papa, Mommy, Lola Jo, Uncle Mike, Tita Emie,
168 Lola Shirley, and Lola Terry – thank you for your unconditional love, patience, and
169 never-ending encouragement. Your strength and support have been the backbone
170 of everything I’ve achieved. Your belief in me has always kept me grounded and

171 motivated, even when I doubted myself. Thank you for the constant prayers,
172 thoughtful check-ins, and warm words that lifted me during the hardest times.
173 Your financial support has also played an instrumental role in helping me pursue
174 this journey, and I am deeply grateful.

175 To my lovely (but sometimes annoying) siblings – Hariette, Vladimir, Michaela,
176 Regina, and Daniel – words can't fully express how much you mean to me. Thank
177 you for being my constant source of joy, laughter, and sanity. Through all the
178 sibling chaos, inside jokes, and heart-to-heart conversations, you've kept me con-
179 nected to the things that truly matter. Whether it was a random meme shared
180 to make me laugh, a quick call just to check in, or simply being there in quiet
181 support, you all reminded me that no matter where I was, I was never truly alone.

182 To my ever-reliable and ever-beautiful Gumamela Girls – Briana, Gliezel, Ari-
183 anne, Sharah, Kai, and Kzlyr – thank you for the love, the chika, and the comfort.
184 You all made even the most stressful days feel manageable. You've truly been my
185 home away from home. Special mention to Briana and Gliezel, my fellow re-
186 searchers, who stood by me not only as friends but as teammates navigating this
187 academic journey together. Your dedication, the leg-torturing stairs we climbed
188 to the hatchery for data gathering, and late-night grind sessions will always be
189 one of my favorite parts of this experience.

190 To the UPV Virgils, thank you for being the silent heroes behind my survival.
191 Your monthly grocery supplies kept me fed, sane, and functioning. I probably
192 owe half of this thesis to the snacks.

193 To the Balay Gumamela dorm staff, thank you for being my guardians during
194 my stay at the dorm. Your support, care, and constant presence made my time

195 away from home more comfortable and secure. I will always be grateful for the
196 way you looked out for me and made me feel safe.

197 To my very cutesy, very demure, very mindful pet dogs, Chewy and Max
198 – thank you for the joy and companionship you've brought into my life. Max,
199 though you are no longer with us, your memory continues to live on in my heart.
200 Chewy, thank you for being my ever-present companion, sharing both the quiet,
201 loud, and the playful moments.

202 To chonky cat Mother Litob, your cuteness and presence in the hatchery dur-
203 ing our data gathering made the 100-step stairs bearable. My tiredness would
204 instantly vanish whenever I petted you.

205 This journey has been a collective effort – held up by the love, wisdom, laugh-
206 ter, and resilience of those around me. I am endlessly grateful for each of you.

Abstract

208 *Tegillarca granosa*, commonly known as blood cockles, is a significant marine bi-
209 valve species due to its nutritional value and economic importance. Accurate sex
210 identification is crucial for maintaining a balanced male-to-female ratio, support-
211 ing sustainable harvesting, and improving resource management. However, macro-
212 scopically identifying sex through shell morphology is challenging, and there are
213 currently no available technologies for non-invasive sex classification. This study
214 explores the use of machine learning and deep learning techniques to classify the
215 sex of blood cockles based on shell measurements (length, width, height, hinge
216 line length, distance between the umbos, and rib count) and images taken from
217 various angles (dorsal, ventral, anterior, posterior, and lateral views). Machine
218 learning analysis using K-Nearest Neighbors (KNN) achieved 64.16% accuracy,
219 64.97% precision, 64.16% recall, and 63.75% F1 Score. Moreover, deep learning
220 using Convolutional Neural Networks (CNN) achieved 71.68% accuracy, 72.52%
221 precision, 69.29% recall, 69.12% F1 Score, and 77.34% AUC score using images
222 captured from the left lateral angle view. These results demonstrate the potential
223 of a non-invasive approach to sex identification, supporting sustainable aquacul-
224 ture practices and offering a baseline for further research using computer vision
225 and machine learning.

226 **Keywords:** deep learning, supervised machine learning, computer vision,
convolutional neural network, blood cockle, sex identifica-
tion, *Tegillarca granosa*

Contents

²²⁷	1	Introduction	1
²²⁹	1.1	Overview	1
²³⁰	1.2	Problem Statement	2
²³¹	1.3	Research Objectives	4
²³²	1.3.1	General Objective	4
²³³	1.3.2	Specific Objectives	4
²³⁴	1.4	Scope and Limitations of the Research	5
²³⁵	1.5	Significance of the Research	6
²³⁶	2	Review of Related Literature	9
²³⁷	2.1	Background on <i>T. granosa</i> and Their Importance	10
²³⁸	2.2	Sex Identification Methods in <i>T. granosa</i>	14

239	2.3 Machine Learning and Deep Learning in Biology	16
240	2.3.1 Deep Learning for Phenotype Classification in Ark Shells .	17
241	2.3.2 Geometric Morphometrics and Machine Learning for Species	
242	Delimitation	18
243	2.3.3 Contour Analysis in Mollusc Shells Using Machine Learning	18
244	2.3.4 Machine Learning for Shape Analysis of Marine Organisms	19
245	2.3.5 Deep Learning for Landmark-Free Morphological Feature	
246	Extraction	21
247	2.3.6 Machine Learning for Sex Differentiation in Abalone . . .	21
248	2.3.7 Machine Learning for Geographical Traceability in Bivalves	23
249	2.4 Limitations on Sex Identification in <i>T. granosa</i>	24
250	2.5 Chapter Summary	26
251	3 Research Methodology	29
252	3.1 Sample Collection	30
253	3.2 Ethical Considerations	31
254	3.3 Creating <i>T. granosa</i> Dataset	32
255	3.4 Morphometric Data Collection	33
256	3.5 Image Acquisition and Data Gathering	34

CONTENTS xv

257	3.6 Hardware and Software Configuration	35
258	3.7 Machine Learning on Morphometric Data	36
259	3.7.1 Data Preprocessing	37
260	3.7.2 Machine Learning Models Training	39
261	3.8 Deep Learning for Morphological Analysis	40
262	3.8.1 Image Preprocessing	40
263	3.8.2 Convolutional Neural Network	42
264	3.8.3 CNN Training	44
265	3.9 Evaluation Metrics	47
266	4 Results and Discussions	51
267	4.1 Machine Learning Analysis	51
268	4.1.1 Data Exploration	52
269	4.1.2 Statistical Analysis	53
270	4.1.3 Feature Importance Analysis	54
271	4.1.4 Performance Evaluation	55
272	4.1.5 Visualizations for Machine Learning	56
273	4.2 Deep Learning Analysis	58

274	4.2.1 Baseline Model	59
275	4.2.2 Comparison of Individual and Combined Angles	60
276	4.2.3 Training Result and Hyperparameter Tuning	61
277	4.2.4 Proposed Model	62
278	4.2.5 Learning Rates and Training Behavior per Fold	63
279	4.2.6 Visualizations for Deep Learning	64
280	4.3 Discussions	69
281	5 Conclusion and Recommendations	73
282	5.1 Conclusion	73
283	5.2 Recommendations	75
284	6 References	77
285	A Code Snippets	85
286	B Resource Persons	91
287	C Data Gathering Documentation	93

²⁸⁸ List of Figures

²⁸⁹	2.1 Diagram of <i>T. granosa</i> 's external anatomy.	12
²⁹⁰	3.1 Female <i>T. granosa</i> shells.	31
²⁹¹	3.2 Male <i>T. granosa</i> shells.	31
²⁹²	3.3 Different views of the <i>T. granosa</i> shell captured.	32
²⁹³	3.4 Linear measurements that were gathered from the shell of <i>T. granosa</i> . .	34
²⁹⁴	3.5 Image acquisition setup for <i>T. granosa</i> samples.	35
²⁹⁵	3.6 Data preprocessing in machine learning pipeline.	37
²⁹⁶	3.7 Diagram of k-fold cross-validation with $k = 5$	40
²⁹⁷	3.8 Shadows removed from male samples at different angles.	41
²⁹⁸	3.9 Shadows removed from female samples at different angles.	42
²⁹⁹	3.10 Architecture of convolutional neural network (CNN).	42
³⁰⁰	3.11 Diagram of stratified k-fold cross-validation with $k=5$	45

301	3.12 On-the-Fly dataset augmentation (OnDAT) techniques.	46
302	4.1 Heatmap of morphometric correlations with <i>T. granosa</i> sex.	52
303	4.2 Feature importance scores using the Kruskal-Wallis test.	54
304	4.3 KNN confusion matrix for <i>T. granosa</i> sex classification.	57
305	4.4 ROC curve with AUC score for KNN.	58
306	4.5 Training and validation accuracy per fold.	65
307	4.6 Average training and validation accuracy across folds.	65
308	4.7 Training and validation loss per fold.	66
309	4.8 Average training and validation loss across folds.	67
310	4.9 Confusion matrix for final model predictions.	68
311	4.10 ROC curve with AUC score for the proposed model.	69
312	A.1 Feature engineering used to create and transform the dataset for 313 machine learning analysis.	85
314	A.2 Feature importance scores derived from the Kruskal-Wallis test to 315 identify the most significant variables.	86
316	A.3 Random undersampling applied in machine learning to address 317 class imbalance.	86

318	A.4 Five-fold cross-validation used to evaluate and tune machine learn-	
319	ing model performance.	87
320	A.5 Resizing images to 256x256 pixels for consistent input dimensions.	87
321	A.6 Processing the images to remove the shadows.	88
322	A.7 Random undersampling applied in deep learning to balance class	
323	distribution in the datasets.	89
324	A.8 On-the-fly data augmentation used to create variety of random	
325	transformation to increase training images.	89
326	A.9 CNN architecture used for training the image dataset.	90
327	C.1 Sex identification through spawning of <i>T. granosa</i>	93
328	C.2 Sex-based separation of <i>T. granosa</i> samples post-spawning.	94
329	C.3 Sex identified female through dissection of <i>T. granosa</i>	94
330	C.4 Sex identified male through dissection of <i>T. granosa</i>	94
331	C.5 Linear measurements of female <i>T. granosa</i>	95
332	C.6 Linear measurements of male <i>T. granosa</i>	95
333	C.7 Captured images of female <i>T. granosa</i>	96
334	C.8 Captured images of male <i>T. granosa</i>	96

List of Tables

336	2.1 Comparison of the methods used in bivalves studies.	27
337	3.1 Architecture of the convolutional neural network used.	44
338	4.1 Mann-Whitney U test results for sex-based feature comparison. .	53
339	4.2 Performance metrics for models with all 13 features.	55
340	4.3 Performance metrics for models with 5 features.	56
341	4.4 Performance metrics for unbalanced vs. balanced datasets (Batch	
342	Size: 16, Epochs: 20).	59
343	4.5 Performance metrics for individual and combined angles (Batch	
344	Size: 16, Epochs: 20).	60
345	4.6 Effect of batch size and epoch values on CNN model performance.	62
346	4.7 Performance metrics for different activation functions (Batch Size:	
347	32, Epochs: 50).	62

348	4.8 Per-fold performance metrics (Batch Size: 32, Epochs: 50, Activation Function: ReLU).	63
350	4.9 Learning rate reductions, early stopping, and best epochs per fold during 5-fold cross-validation.	64
351		

³⁵² Chapter 1

³⁵³ Introduction

³⁵⁴ 1.1 Overview

³⁵⁵ The Philippines is a global center of marine biodiversity and has established aqua-
³⁵⁶ culture as a significant contributor to total fishery production (Aypa & Baconguis,
³⁵⁷ 2000; BFAR, 2019). The country produces over 4 million tonnes of seafood annu-
³⁵⁸ ally and is the 11th largest seafood producer in the world. Aquaculture is deeply
³⁵⁹ integrated into Filipinos' livelihoods, encompassing fish cultivation and the pro-
³⁶⁰ duction of various aquatic species, including bivalves. Among these, blood cockles
³⁶¹ (*Tegillarca granosa*) hold considerable economic and environmental significance,
³⁶² making it essential to ensure sustainable production and population balance.

³⁶³ Maintaining a balanced male-to-female ratio of blood cockles is crucial to prevent
³⁶⁴ overharvesting and ensure sustainability. An imbalanced ratio can lead to over-
³⁶⁵ exploitation and negatively impact the population's viability. However, there is

³⁶⁶ limited literature on *T. granosa* that provides a thorough understanding of its
³⁶⁷ sex-determining mechanisms, particularly regarding sexual dimorphism based on
³⁶⁸ morphometric and morphological characteristics (Breton, Capt, Guerra, & Stew-
³⁶⁹ art, 2017).

³⁷⁰ Currently, sex determination methods for blood cockles are invasive, including
³⁷¹ dissection and histological examinations, which often result in the death of the
³⁷² species. While there is growing literature on sex identification in aquaculture
³⁷³ commodities using machine learning and deep learning, there is a notable scarcity
³⁷⁴ of research specific to *T. granosa* (Miranda & Ferriols, 2023).

³⁷⁵ This study aims to provide a detailed baseline analysis of blood cockles by lever-
³⁷⁶ aging their morphometric and morphological characteristics. Sexual dimorphism
³⁷⁷ in bivalves is often subtle and challenging to establish macroscopically (Karapunar,
³⁷⁸ Werner, Fürsich, & Nützel, 2021). However, by integrating machine learning and
³⁷⁹ deep learning, the study seeks to identify distinct features that may indicate sexual
³⁸⁰ dimorphism between male and female blood cockles.

³⁸¹ 1.2 Problem Statement

³⁸² Identifying the sex of *Tegillarca granosa* is important for promoting sustainable
³⁸³ aquaculture and biodiversity by maintaining a balanced male-to-female ratio. A
³⁸⁴ balanced ratio helps prevent overharvesting. Although sex identification is crucial
³⁸⁵ for blood cockle population management and sustainable aquaculture, there is a
³⁸⁶ notable lack of research on creating non-invasive methods for determining the sex
³⁸⁷ of *T. granosa*. Many recent studies and approaches rely on invasive methods like

388 dissection or histological analysis, which are impractical for large-scale aquaculture
389 operations focused on conservation.

390 Current methods for determining the sex of *T. granosa* are invasive and involve
391 dissection, which requires cutting open the shell to visually inspect the gonads
392 (Erica, 2018). This procedure can cause harm to the specimens and frequently
393 leads to their death. Another method is histological examination, where tissue
394 samples are analyzed under a microscope (May, Maung, Phy, & Tun, 2021). Both
395 approaches are labor-intensive and time-consuming, and can pose risks to popula-
396 tion management, particularly when maintaining a balanced sex ratio for breeding
397 programs is essential. Moreover, these invasive methods require specialized tech-
398 nical skills for accurate execution. Resource-limited aquaculture operations face
399 significant challenges in accessing the necessary laboratory equipment, such as
400 microscopes and staining tools, complicating the process.

401 A less invasive approach employed by aquaculturists involves monitor spawning
402 behavior, where individuals are separated and stimulated to reproduce in order
403 to determine their sex through the release of gametes (Miranda & Ferriols, 2023).
404 Although this method is indeed less invasive than dissection, it still induces stress
405 in blood cockles and may not be completely effective for fast identification in large
406 populations.

407 Given the limitations of both invasive and less invasive methods, there is a clear
408 need for a more advanced approach. An alternative, non-invasive method involv-
409 ing machine and deep learning technologies could address these issues by provid-
410 ing a fast, accurate, and effective solution without harming or stressing the blood
411 cockles.

412 1.3 Research Objectives

413 1.3.1 General Objective

414 The general objective of this study is to develop a non-invasive method for iden-
415 tifying the sex of *Tegillarca granosa* using machine learning and deep learning
416 technologies. This method aims to provide accurate and streamlined sex iden-
417 tification without causing harm to the specimens, thus supporting sustainable
418 aquaculture practices.

419 1.3.2 Specific Objectives

420 To achieve the overall general objective of developing a non-invasive sex identifi-
421 cation of *T. granosa* using machine learning and deep learning technologies, the
422 following specific objectives have been established:

- 423 1. to collect and organize a comprehensive dataset of *T. granosa*, which will
424 include linear measurements and images captured from different camera an-
425 gles that will serve as the basis for training and evaluating the machine
426 learning and deep learning models,
- 427 2. to develop and implement machine learning and deep learning models that
428 can classify the sex of *T. granosa* based on the collected linear measurements
429 and images of different camera angles of the sample, and determine the best
430 performing models, and
- 431 3. to evaluate the model using performance metrics such as accuracy, precision,

432 recall, F1 Score, and AUC-ROC score for deep learning, and improve it by
433 performing hyperparameter optimization.

434 1.4 Scope and Limitations of the Research

435 This study is conducted alongside the ongoing research by the UPV DOST-
436 PCAARRD, titled "Establishment of the Center for Mollusc Research and De-
437 velopment: Development of Spawning and Hatchery Techniques for the Blood
438 Cockle (*Anadara granosa*) for Sustainable Aquaculture." The ongoing research
439 primarily involves the rearing of *Tegillarca granosa* from spat to larvae, feeding
440 experiments, stocking density evaluations, substrate selection, and settlement rate
441 assessments.

442 In contrast, this study mainly focused on developing a non-invasive method for
443 identifying the sex of *T. granosa* using machine learning and deep learning tech-
444 nologies. The goal is to provide an accurate and efficient means of sex identifica-
445 tion without causing harm to the samples, contributing to sustainable aquaculture
446 practices.

447 The researchers worked with 271 blood cockles that had been sex-identified and
448 taken from Panay Island, specifically sourced from Zarraga Iloilo and Ivisan Capiz.
449 These samples, divided between 144 males and 127 females, were obtained through
450 induced spawning via temperature shock and dissection. Data collection was lim-
451 ited to the spawned stage among the five gonadal stages - immature, developing,
452 mature, spawning, and spent stages. The other stages were not preferable due to
453 indistinguishable gonads and their inability to undergo induced spawning (May

⁴⁵⁴ et al., 2021). Thus, the researchers only focused on the samples undergoing the
⁴⁵⁵ spawned stage.

⁴⁵⁶ During the data collection, the researchers personally gathered linear measure-
⁴⁵⁷ ments, including length, width, height, rib count, hinge line length, and distance
⁴⁵⁸ between the umbos through the vernier caliper. The data gathering process was
⁴⁵⁹ supervised by the University Research Associates from the Institute of Aquacul-
⁴⁶⁰ ture, College of Fisheries and Ocean Sciences. Aside from linear measurements,
⁴⁶¹ images were taken from six different angles. The process of linear measurements
⁴⁶² and image collection were non-invasive, considering the blood cockle-built ability
⁴⁶³ to survive in low oxygen environments and naturally inhabit intertidal mudflats
⁴⁶⁴ (Zhan & Bao, 2022).

⁴⁶⁵ The method developed in this study is specific to *T. granosa* and may not apply to
⁴⁶⁶ other bivalve species. The model was trained exclusively for *T. granosa*'s morpho-
⁴⁶⁷ metric and morphological features, which may not be consistent and applicable
⁴⁶⁸ across other shellfish species.

⁴⁶⁹ 1.5 Significance of the Research

⁴⁷⁰ This study will give us a significant advancement in non-invasive sex identification
⁴⁷¹ methods in *Tegillarca granosa*, providing innovative solutions that has the poten-
⁴⁷² tial to address the challenges in identifying sex and reshape sustainable approaches
⁴⁷³ to aquaculture. The significance of this study extends to the following:

⁴⁷⁴ *Research Institution.* The result of this study focusing on the sex-identification

475 mechanism of bivalves, specifically *T. granosa*, will provide valuable insights into
476 universities and research centers that focus on fisheries and coastal management,
477 such as the UPV Institute of Aquaculture, that aim to develop sustainable devel-
478 opment and suitable culture techniques.

479 *Fishermen.* By developing a non-invasive method in sex identification, this study
480 can help long-term harvest efficiency and maintain the ratio of the harvest which
481 can help prevent exploitation of the *T. granosa*.

482 *Coastal Communities.* The result of this study would be beneficial for the coastal
483 communities that are reliant on their source of income with aquaculture com-
484 modities like blood cockles. Maintaining the diversity and aspect ratio of male
485 and female may increase the market value of blood cockle production since cockle
486 aquaculture faces significant obstacles worldwide due to the fluctuating seed sup-
487 plies and scarcity of broodstock from the wild.

488 *Future Researchers.* The result of this study would serve as the basis for studies
489 that involve sex identification in bivalves such as *T. granosa*. Some technologies
490 are yet to be explored in machine learning and deep learning technologies that
491 can lead to higher accuracy and distinguish the presence of sexual dimorphism in
492 the *T. granosa*.

⁴⁹³ Chapter 2

⁴⁹⁴ Review of Related Literature

⁴⁹⁵ Aquaculture is the fastest-growing industry in animal food production and has
⁴⁹⁶ great potential as a sustainable solution to global food security, nutrition, and
⁴⁹⁷ development (*FAO 2024 Report: Sustainable Aquatic Food Systems Important*
⁴⁹⁸ *for Global Food Security – European Fishmeal*, 2024). Aquaculture is deeply in-
⁴⁹⁹ tegrated into the livelihoods of Filipinos, not only through fish cultivation but
⁵⁰⁰ also through the production of other aquatic species, including mollusks, oysters,
⁵⁰¹ clams, scallops, and mussels (Breton et al., 2017). Mollusks, particularly blood
⁵⁰² clams *Tegillarca granosa*, have economic and environmental significance. It has
⁵⁰³ been a collective effort to maintain an ideal male-to-female ratio to avoid overhar-
⁵⁰⁴ vesting and maintain the optimal ratio to preserve the population and production
⁵⁰⁵ of the blood cockles.

⁵⁰⁶ The members of the Arcidae Family, including *T. granosa* are important sources
⁵⁰⁷ of food and livelihood. Cockle aquaculture meets rising demands, however, it
⁵⁰⁸ faces significant challenges due to fluctuating seed supplies (Miranda & Ferriols,

509 2023). To solve the problem, researchers exert a considerable amount of effort,
510 developing a broader understanding of bivalves, including their sex-determining
511 mechanism, due to their notable importance in terms of diversity, environmental
512 benefits, and economic and market importance (Breton et al., 2017). Despite the
513 promising idea of identifying sex, there is limited research reported in terms of
514 sexual dimorphism, making it harder to distinguish through its morphological and
515 morphometric characteristics.

516 By addressing the challenges in the sex identification of *T. granosa*, it would be
517 able to address one problem at a time. Currently, there are no recent documented
518 publications that integrate machine learning and deep learning in characterizing
519 sexual dimorphism, reducing complexity, variability in sex determination, and
520 differentiation mechanisms in bivalves, including *T. granosa* specifically.

521 **2.1 Background on *T. granosa* and Their Im- 522 portance**

523 *Tegillarca granosa* (Linnaeus, 1758) is also known as blood cockles or blood clam.
524 In the Philippines, it is known locally as Litob and Bakalan, a marine bivalve
525 species from the family Arcidae. Litob is widely distributed in the world including
526 Southeast Asia. They can be found in the intertidal mudflats adjacent to the
527 mangrove forest (Srisunont, Nobpakhun, Yamalee, & Srisunont, 2020). With
528 the intertidal mudflat as *T. granosa*'s habitat, they experience severe hypoxia
529 or low oxygen levels in the blood tissues during the tidal cycle. The blood clams
530 exhibit a unique red-blood phenotype where it serves two purposes the hemocyte

531 carries oxygen around the body and strengthens immune defenses. In addition,
532 it possesses a unique ability to absorb oxygen at similar rates in water and air
533 (Zhan & Bao, 2022).

534 *T. granosa* shell (refer to Figure 2.1) is medium-sized, fairly thick, ovate, and
535 convex, with both valves being equal in size but asymmetrical from the hinge. The
536 top edge of the dorsal margin is straight, while the front is rounded and slopes
537 downward, with its back being obliquely rounded with a concave bottom edge.
538 It has a narrow diamond-shaped ligament near the hinge with 3-4 dark chevron
539 markings, although some may be incomplete. The shell's outer layer, or the
540 periostracum, is smooth and brown with a straight hinge line and 40-68 fine short
541 teeth arranged in a straight line. The beak, or prosogyrate, curves forward, with
542 the shell having 18–21 raised ribs with blunt nodules and spaces between them.
543 The inner shell is white with crenulations along the valves' ventral, anterior, and
544 posterior margins. The posterior adductor scar is elongated and squarish, while
545 the anterior adductor scar is similar but smaller in size. The mantle covering the
546 bulk of *T. granosa*'s visceral mass is thin but the edges are thick and muscular.
547 It bears the impression of the crenulated shell edges. Their foot is large with a
548 ventral grove with no byssus or thread-like attachment. The *T. granosa*'s soft
549 body is blood red (Narasimham, 1988).

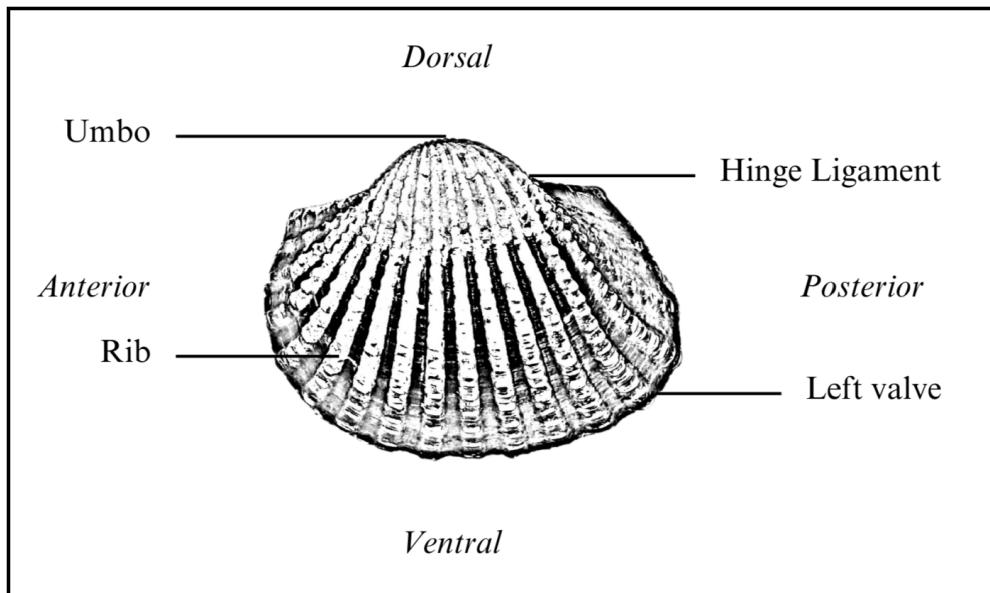


Figure 2.1: Diagram of *T. granosa*'s external anatomy.

550 *T. granosa* is one of the most well-known marine bivalves given that they are a
551 protein-rich food, known for their rich flavor, substantial nutritional benefits, a
552 good source of vitamins, low in fat, and contain a considerable amount of iron,
553 important in combating anemia (Zha et al., 2022). Blood cockles were collected
554 by locals inhabiting the brackish mudflats during the low tides for consumption
555 and sold in the market as a source of livelihood (Miranda & Ferriols, 2023). *T.*
556 *granosa* is not only valuable for its market and food purposes but also facilitates
557 an important role in marine ecosystems as a food source for various organisms
558 like wading birds, intertidal-feeding fish, and crustaceans such as shore crabs and
559 shrimp (Burdon, Callaway, Elliott, Smith, & Wither, 2014). Blood cockles can act
560 as sentinel species and a bioindicator of marine pollutants such as heavy metals
561 (Ishak, Mohamad, Soo, & Hamid, 2016) and polycyclic aromatic hydrocarbons
562 (PAHs) (Sany et al., 2014). Additionally, cockle shells can be utilized to create a
563 cost-effective catalyst for biodiesel production by providing calcium oxide (Boey,

564 Maniam, Hamid, & Ali, 2011).

565 Determining the sex of bivalves is important for three reasons: diversity, envi-
566 ronmental benefits, and economic significance (Breton et al., 2010). Firstly, with
567 the estimated 25,000 living species under class Bivalvia, it would be a suitable
568 resource to develop a broader understanding of their evolution of the sex and sex
569 determination mechanism (Breton et al., 2010). Second, studying sex determi-
570 nation is important since bivalves are utilized as bioindicators of environmental
571 health. This would pave the way for understanding bivalves' life cycle and popula-
572 tion dynamics in determining different factors that affect them (Campos, Tedesco,
573 Vasconcelos, & Cristobal, 2012). Thirdly, the immediate and practical reason to
574 unveil the sex determination mechanism is the economic and nutritional impor-
575 tance of bivalves as a large population of people relies on fish and shellfish as
576 sources of food and nutrition (Naylor et al., 2000). Additionally, male and female
577 aquaculture commodities have different growth and economic values. Male Nile
578 tilapia, for example, grow faster and have lower feed conversion rates than females,
579 female Kuruma prawns (*Penaeus japonicus*) are generally larger than males at the
580 time of harvest (Budd, Banh, Domingos, & Jerry, 2015).

581 Clearly, much more work is required to understand the mechanisms underlying
582 sexual dimorphism in bivalves, specifically *T. granosa*. Just like the other aqua-
583 culture commodities, sex affects not just reproduction but it can also affect market
584 preference and underlying economic value, making the determination of sex im-
585 portant for meeting consumer demands. These are the increasing significance of
586 the *T. granosa* despite the lack of reviewed articles in the Philippines.

587 2.2 Sex Identification Methods in *T. granosa*

588 The current sex identification methods in *Tegillarca granosa* range from invasive
589 histological techniques to less invasive methodologies like temperature-induced
590 spawning. Each approach comes with its pros and cons regarding accuracy, feasi-
591 bility, and impact on natural populations.

592 Induced spawning and larval rearing are considered the less invasive techniques
593 used to study *T. granosa*. In the Philippines, limited research has been done
594 on the *T. granosa* (Linnaeus, 1758), and this study, titled Initial Attempts on
595 Spawning and Larval Rearing of the Blood Cockle, *T. granosa* in the Philippines,
596 was conducted by Miranda and Ferriols (2023). The researchers conducted ex-
597 periments on induced spawning and larval rearing, discovering that the eggs of
598 female *T. granosa* were salmon pink, while the sperm released by males looked
599 milky. After spawning, the researchers successfully generated 6,531,000 fertilized
600 eggs.

601 The researchers highlighted the importance of *T. granosa* and other anadarinids as
602 a food source established worldwide, especially in Malaysia and Korea. However,
603 in the Philippines, the bivalve aquaculture of the clam species is still limited. The
604 experiment, which focused on the culture and rearing of *T. granosa*, was attempted
605 by subjecting the wild broodstocks to a series of temperature fluctuations to
606 induce the spawning of gametes. This is currently the most natural and least
607 invasive sex identification method for bivalves (Aji, 2011). The study of Miranda
608 and Ferriols aimed to pave the way for the sustainable production of *T. granosa*
609 seeds for aquaculture and stock enhancement, despite the scarcity of documented
610 hatchery culture of *T. granosa* from larvae to adults in the Philippines.

611 On the other hand, invasive techniques such as histological analysis offer a more
612 thorough but harmful method for determining the sex of *T. granosa*. A study on
613 the spawning period of blood cockle *T. granosa* (Linnaeus, 1758) in the Myeik
614 coastal area examined 240 blood cockle samples for sex and gonad maturity stages
615 using histological examination, with shell lengths ranging from 26–35 mm and
616 shell weights from 8.1–33 g. For histological analysis, the whole soft tissues were
617 removed from the shell and the flesh containing most parts of the gonads was fixed
618 in formalin, dehydrated in an upgraded series of ethanol, and cleared in xylene.
619 This invasive method allows for precise identification of the gonadal maturation
620 stages based on cellular and structural changes in the gonads.

621 The classification of the gonad stages used was by Yurimoto et al. (2014). There
622 are five maturation stages of gonadal development: immature (Stage I), devel-
623 oping (Stage II), mature (Stage III), spawning (Stage IV), and spent (Stage V)
624 stages. The sex of the *T. granosa* was confirmed by the color of the gonad and
625 by conducting a histological examination of the gonads. During the immature
626 stage, sex determination was indistinguishable due to the difficulties of observing
627 the germ cells. In the developing stage, the spermatocytes and a few spermatids
628 can be seen for males, and immature oocytes are attached to the tube wall for
629 the female. In the mature stage, the follicles are full of spermatozoa with their
630 tails pointing towards the center of the tube for the male, and the female is full
631 of mature oocytes that are irregular or polygonal in shape with the oval nucleus.
632 Upon reaching spawning, some spermatozoa are released, causing the empty space
633 in the follicle wall for males and females. There is a decrease in the number of
634 mature oocytes and it exhibits nuclear disappearance due to the breakdown of
635 the germinal vesicle. Lastly, the spent stage is where the genital tube is deformed

636 and devoid of spermatocytes which have completely spawned. In the female, the
637 genital tube is deformed and degenerated, making it empty. The morphology of
638 the cockle gonad shows that the area of the gonad increases according to the in-
639 creased levels of gonad maturity. The coloration of the gonad tissue layer in the
640 blood cockle varies from orange-red to pale orange in females and from white to
641 grayish-white in males for different maturity stages (May et al., 2021).

642 Although the histological examination is the most reliable method for obtain-
643 ing accurate information on the reproductive biology and sex determination of
644 *T. granosa*, it has limitations. Given its invasive nature, this approach requires
645 the dissection and destruction of specimens, making it unsuitable for continuous
646 monitoring and conservation efforts. Moreover, the current understanding of sex
647 determination in bivalves and mollusks is poor, and no chromosomes that can
648 be differentiated based on their morphology have been discovered (Afiati, 2007).
649 There exists a study that can provide insight into the sex-determining factor in
650 bivalves but *N. schoberi* is more difficult to analyze concerning potential sexual
651 dimorphism. Thickening the edges of the shell increases its inflation, which means
652 the shell can hold more space inside. This extra space helps protandrous females
653 accommodate more eggs.

654 2.3 Machine Learning and Deep Learning in Bi- 655 ology

656 Machine learning has the potential to improve the quality of life of human beings
657 and has a wide range of applications in terms of research and development. The

term machine learning refers to the invention and algorithm evaluation that enables pattern recognition, classification, and prediction based on models generated from available data (Tarcă, Carey, Chen, Romero, & Drăghici, 2007). The study of machine learning methods has advanced in the last several years, including biological studies. In biological studies, machine learning has been used for discovery and prediction. This section will explore existing machine learning studies that are applied in biological sciences, highlighting the identification of sex in shells, bivalves, and mollusks.

2.3.1 Deep Learning for Phenotype Classification in Ark Shells

In the study by Kim et al. (2024), the researchers utilized three (3) convolutional neural network (CNN) models: the Visual Geometry Group Network (VGGnet), the Inception Residual Network (ResNet), and the SqueezeNet. These deep learning models are utilized for the ark shells, namely *Anadara kagoshimensis*, *Tegillarca granosa*, and *Anadara broughtonii*, to identify the phenotype classification.

The researchers classified the ark shells based on radial rib count where they investigated the difference in the number of radial ribs between three species and were counted. Their CNN-based model that classifies images of three ark shells can provide a theoretical basis for bivalve classification and enable the tracking of the entire production process of ark shells from catching to selling with the support of big data, which is useful for improving food safety, production efficiency, and economic benefits (Kim, Yang, Cha, Jung, & Kim, 2024).

681 **2.3.2 Geometric Morphometrics and Machine Learning for**
682 **Species Delimitation**

683 In *Geometric morphometrics and machine learning challenge currently accepted*
684 *species limits of the land snail Placostylus (Pulmonata: Bothriembryontidae) on*
685 *the Isle of Pines, New Caledonia*, the shell size was quantified using centroid size
686 from the Procrustes analysis, and both the shape and size information were used in
687 training the machine learning model. Their study concluded that the researchers
688 support utilizing both methods: supervised and unsupervised machine learning,
689 rather than choosing either of them individually. In general, their research con-
690 tributes to the growing number of studies that have combined geometric morpho-
691 metrics with the aid of machine learning, which is helpful in biological innovation
692 and breakthrough (Quenu, Trewick, Brescia, & Morgan-Richards, 2020).

693 **2.3.3 Contour Analysis in Mollusc Shells Using Machine**
694 **Learning**

695 Tuset et al. (2020), in their study, *Recognising mollusc shell contours with enlarged*
696 *spines: Wavelet vs Elliptic Fourier analyses*, mentioned that gastropod shells have
697 large spines and sharp shapes that differ based on environmental, taxonomic, and
698 evolutionary influences. The researchers stated that classic morphometric meth-
699 ods may not accurately depict morphological features of the shell, especially when
700 using the angular decomposition of the contour. The current research examined
701 and compared the robustness of the contour analysis using wavelet transformed
702 and Elliptic Fourier descriptors for gastropod shells with enlarged spines. For

703 that, the researchers analyzed two geographically and ecologically separated pop-
704ulations of *Bolinus brandaris* from the NW Mediterranean Sea. Results showed
705that contour analysis of gastropod shells with enlarged spines can be analyzed
706using both methodologies, but the wavelet analysis provided better local discrim-
707ination. From an ecological perspective, shells with various sizes of spines in both
708areas indicate the broad adaptability of the species.

709 **2.3.4 Machine Learning for Shape Analysis of Marine Or- 710 ganisms**

711 In the study of Lishchenko and Jones (2021), titled *Application of Shape Analyses*
712 *to Recording Structures of Marine Organisms for Stock Discrimination and Taxo-*
713 *nomic Purposes*, they utilized geometric morphometrics (GM) as an approach to
714 the traditional method of collecting linear measurements with the application of
715 multivariate statistical methods and outline analysis in recording the structures
716 of marine organisms. The main taxonomic categories (mollusks, teleost fish, and
717 elasmobranchs) with their hard bodies have been used as an indication of age and
718 a determinable time-scale and structure continue to go through life (Arkhipkin,
719 2005; Kerr & Campana, 2014). This study has explored variations in the mor-
720 phometry of recording structures in stock discrimination and systematics. The
721 researchers utilized the principal component analysis rather than the traditional
722 approach, which helps simplify the data without losing important information.
723 They utilized landmark-based geometric morphometrics, which has three differ-
724 ent types, namely: discrete juxtaposition of tissue, maxima or curvature, or other
725 morphogenetic processes, and lastly, the extremal points are constructed land-

726 marks.

727 Generalized Procrustes Analysis (GPA) is a common superimposition technique in
728 landmark-based geometric morphometrics that aligns landmarks via translation,
729 scaling, and rotation to eliminate non-shape deviations (Zelditch, Swiderski, &
730 Sheets, 2004). However, there is a limit to the amount of smooth areas that may
731 be captured, and it is possible to overlook significant shape details. Utilization
732 of the semi-landmarks enhanced the shape description (Adams, Rohlf, & Slice,
733 2004). The researchers observed that using an outline-based approach would be
734 more effective than using a landmark-based approach.

735 Another approach is the Fourier analysis which is a curve-fitting approach com-
736 monly used due to its well-known mathematical background and how general
737 functions can be decomposed into trigonometric or exponential functions with
738 definite frequencies. It has two main approaches, namely: Polar Transform (PT)
739 in which it expresses the outline using equally spaced radii, and Elliptical Fourier
740 Analysis (EFA) which separately analyzes the x and y coordinates of the shape.
741 The PT works for simple rounded outlines and has the tendency to miss details
742 in more complex shapes, unlike the EFA which can handle complex, convoluted
743 outlines (Zahn & Roskies, 1972; Doering & Ludwig, 1990; Ponton, 2006). Many
744 researchers view EFA as the most effective Fourier method for providing a compre-
745 hensive and detailed description of recording structures (Mérigot, Letourneau, &
746 Lecomte-Finiger, 2007; Ferguson, Ward, & Gillanders, 2011; Leguá, Plaza, Pérez,
747 & Arkhipkin, 2013; Mahé et al., 2016).

748 Landmark-based methods used in the study showed that there are detectable
749 differences between male and female octopuses. However, the accuracy of deter-

750 mining sex based on these differences was low, similar to the results obtained
751 with traditional morphometric techniques. The study involved a relatively small
752 sample size of 160 individuals, and the structure being analyzed (the stylet, or
753 internalized shell) varies significantly between individuals. Although the results
754 aligned with findings from other studies that attempted to identify gender differ-
755 ences in cephalopods, the researchers concluded that the approach might not be
756 accurate enough for reliable sex determination.

757 **2.3.5 Deep Learning for Landmark-Free Morphological Fea-
758 ture Extraction**

759 In another study, *a deep learning approach for morphological feature extraction*
760 *based on variational auto-encoder: an application to mandible shape*, the Morpho-
761 VAE machine learning approach was used to conduct a landmark-free shape ana-
762 lysis. Morpho-Vae reduces dimensions by concentrating on morphological features
763 that distinguish data with different labels using an image-based deep learning
764 framework that combines unsupervised and supervised machine learning. After
765 utilizing the method in primate mandible images, the morphological features re-
766 veal the characteristics to which family they belonged. Based on the result, the
767 method applied provides a versatile and promising tool for evaluating a wide range
768 of image data of biological shapes including those missing segments.

769 **2.3.6 Machine Learning for Sex Differentiation in Abalone**

770 In the study, *Towards Abalone Differentiation Through Machine Learning*, re-
771 searchers identified a problem in abalone farming which is having to identify the

772 sex of abalone to apply measures for its growth or preservation. The researchers
773 classified abalone sex using machine learning. Researchers trained the machine
774 to classify different types of classes which are male, female, and immature. The
775 results demonstrated the effectiveness of utilizing linear classifiers for this task.

776 Similarly, in the study, *Data scaling performance on various machine learning*
777 *algorithms to identify abalone sex*, the researchers of the University of India (2022)
778 focused on the data scaling performance of various machine learning algorithms to
779 identify the abalone sex, specifically using min-max normalization and zero-mean
780 standardization. The different machine learning algorithms are the Supervised
781 Vector Machine (SVM), Random Forest, Naive Bayesian, and Decision Tree. Their
782 study aims to utilize machine learning in terms of identifying the trends and
783 distribution patterns in the abalone dataset. Eight features of the abalone dataset
784 (length, diameter, height, whole weight, shucked weight, viscera weight, shell
785 weight, ring) were used to determine the three sexes of Abalone. Their data has
786 been grouped based on sex which are Female, Male, and Infant. They utilized
787 the Synthetic Minority Oversampling Technique (SMOTE) in data balancing for
788 the preprocessing of the data. Followed by data scaling or normalization where
789 it converts numeric values in a data set to a general scale without distorting
790 differences in the range of values. Then they classified by splitting the data into
791 training and testing sets (Arifin, Ariawan, Rosalia, Lukman, & Tufailah, 2021).

792 The study found that Naive Bayes consistently performed better than other algo-
793 rithms. However, when applied to both min-max and zero-mean normalization,
794 the average accuracies of the algorithms were as follows: Random Forest (62.37%),
795 SVM with RBF kernel (59.49%), Decision Tree (57.20%), SVM with linear kernel
796 (56.59%), and Naive Bayes (53.39%). Despite the performance decrease with nor-

797 malization, Random Forest achieved the highest overall metrics, including an av-
798 erage balanced accuracy of 787%, sensitivity of 66.43%, and specificity of 83.31%.
799 Liu et al. concluded that Random Forest is highly accurate because it can handle
800 large, complex datasets, run processes in parallel using multiple trees, and select
801 the most relevant features to enhance model performance (Arifin et al., 2021).

802 **2.3.7 Machine Learning for Geographical Traceability in
803 Bivalves**

804 In the study, *BivalveNet: A hybrid deep neural network for common cockle (*Ceras-**

805 *toderma edule*) geographical traceability based on shell image analysis, the re-
806 searchers incorporated computer vision and machine learning technologies for an
807 efficient determination of blood cockle harvesting origin based on the shell geomet-
808 ric and morphometric analysis. It aims to improve the traceability methodologies
809 in these organisms and its potential as a reliable traceability tool. Thirty *Cerasto-*
810 *derma edule* samples were collected along the five locations on the Atlantic West
811 and South Portuguese coast with individual images processed using lazy snapping
812 segmentation, spectro-textural-morphological phenotype extraction, and feature
813 selection through hybrid Principal Component Analysis and Neighborhood Com-
814 ponent Analysis (Concepcion, Guillermo, Tanner, Fonseca, & Duarte, 2023).

815 The researchers developed a non-invasive image-based traceability technique, an
816 alternative to the chemical and biochemical analysis of the bivalves. It was able
817 to incorporate machine learning methods to promote lesser human intervention.
818 The researchers discovered that BivalveNet emerged as the superior model for
819 bivalves with 96.91% accuracy which is comparable to the accuracy of the de-

820 structive methods with 97% and 97.2% accuracy rates. The result of the study
821 aided the researchers in concluding that there is a possibility of on-site evalua-
822 tion of the bivalve through the implementation of a mobile app that would allow
823 the public and official entities to obtain information regarding the provenance of
824 seafood products' traceability because of its non-invasive and image-based aspects
825 (Concepcion et al., 2023).

826 *T. granosa* is known for having no sexual dimorphism. However, through several
827 related studies, the researchers can apply how family shells of *T. granosa* have
828 been identified based on its morphological and morphometric characteristics and
829 the methods used in machine learning in identifying its sex.

830 **2.4 Limitations on Sex Identification in *T. gra-***

831 ***nosa***

832 To date, no distinction has been made between the male and female *T. granosa*
833 in sexing methodology. In cockle aquaculture without clearly apparent sexual
834 dimorphism, sexing can be performed using invasive methods such as chemical
835 stimulation, dissection, and gonad-stripping. Induced spawning, specifically tem-
836 perature shock, is the most natural and least invasive method for bivalves (Aji,
837 2011). However, the method (Wong & Lim, 2018) of immersing cockles in water
838 from hot to cold with a specific temperature requires deliberate and careful ma-
839 nipulation of the temperature over a specific period and would require constant
840 management and monitoring.

841 Recent studies involved non-invasive methods, with a specific emphasis on mor-
842 phological characteristics as indicators of sex differentiation. However, Tatsuya
843 Yurimoto et al. (2014) stated that the existing methods for determining the sex of
844 bivalves and mollusks in general are somewhat limited (Afiati, 2007). At present,
845 there is no recorded evidence of sexual dimorphism in *T. granosa*. Gonochoristic
846 is the classification given to *T. granosa* (Lee, 1997). However, Lee et al. (2012)
847 reported that the sex ratio varied with shell length, suggesting that sex might
848 alter.

849 Hermaphrodites can exhibit either sequential (asynchronous) or simultaneous (syn-
850 chronous or functional) characteristics. Sequential hermaphrodites switch genders
851 after being male or female for one or multiple yearly cycles. (Heller, 1993; Gosling,
852 2004; Collin, 2013). Sex change and consecutive hermaphroditism have been ob-
853 served in different bivalve species, including Ostreidae, Pectinidae, Veneridae,
854 and Patellidae. However, macroscopically differentiating bivalve sex is challeng-
855 ing. The only way it may be identified is through histological analysis of gonad
856 remains but to do so there is an act of killing the organism (Coe, 1943; Gosling,
857 2004). Verification of sex change in bivalves to classify whether male or female
858 while they are alive is challenging since they need to be re-confirmed and re-
859 evaluated to be the same individual after a year.

860 Lee et al. (2012) found out that *T. granosa*, a species in Arcidae, has been dis-
861 covered to be a sequential hermaphrodite, with the sex ratio changing with an
862 increase in the shell size. In bivalves, sex changes usually happen when the gonad
863 is not differentiated between spawning seasons (Thompson, Newell, Kennedy, &
864 Mann, 1996). But in *T. granosa*, after the spawning season, sex changes during
865 its inactive phase. Results showed a 15.1% sex change ratio, with males having

866 a higher sex change ratio (21.2%) than females (6.2%). The 1+ year class had a
867 higher ratio (17.8%) than the 2+ year class (12.1%). Thus, this study indicates
868 that *T. granosa* is a sequential hermaphrodite. The results of the study demon-
869 strated that the bivalve's age affects the sex ratio and degree of sex change, but
870 additional in-depth investigation is required to determine the role that genetic
871 and environmental factors play in these changes.

872 No literature in the study of mollusks specifically addresses the machine learning
873 and deep learning technologies used to determine the sex of *T. granosa* bivalves in
874 various models. Nevertheless, various techniques such as shape analysis, morpho-
875 metric analysis, Wavelet, and Fourier analysis, as well as different deep learning
876 models like VGNet, ResNet, and SqueezeNet in CNN networks, are utilized for
877 phenotype classification, while different machine learning algorithms could serve
878 as the foundation for this research project.

879 2.5 Chapter Summary

880 This section summarizes the methodologies and problems discussed in other lit-
881 erature encompassing machine learning, deep learning, and other related bivalve
882 studies.

Author	Technology / Method Used	Description of Problem	Pros	Cons
D. V. Miranda and V. M. E. N. Ferriols	Temperature shock	No recent studies are available on the production and rearing of <i>T. granosa</i> in the Philippines.	Employed less invasive techniques which minimize the stress in <i>T. granosa</i> and can lead to better survival rates.	Time-consuming as the entire process from fertilization to the spat stage took 120 days.
Karapunar, Baran and Werner, W. and Fürsich, F. T. and Nützel, A.	Morphometric analysis, microscope imaging, principal component analysis (PCA), and Fourier shape analysis	To address the observed shell dimorphism in the Early Jurassic bivalve <i>Nicanella rakoveci</i> , namely the presence or lack of crenulations on the ventral shell margin, and whether these variations represent sexual dimorphism and sequential hermaphroditism.	The methods used reveal significant morphological differences with regard to sexual dimorphism.	There could be misinterpretation of the shape differences of bivalves due to the constraints and resolution of technologies used.
K. May and C. Maung and E. Phyus and N. Tun	Histological examination	The need to understand the reproductive period of <i>T. granosa</i> in Myeik to ensure sustainable aquaculture and to prevent overexploitation.	Method used allows for accurate sex identification based on the histological characteristics and color of the gonads.	Invasive technique used to determine the sex of <i>T. granosa</i> through gonad histological analysis.
E. Kim and S.-M. Yang and J.-E. Cha and D.-H. Jung and H.-Y. Kim	Convolutional neural network (CNN) models, VGGNet, Inception-ResNet, SqueezeNet	Traditional methods of recognizing and classifying ark shell species based on shell traits are time-consuming and inaccurate.	Automated classification of the three ark shells using a deep learning model obtained an accuracy of 92.4%.	Challenges may arise with certain ark shells that share similar morphology.
Mathieu Quemu and S. A. Trewick and F. Brescia and M. Morgan-Richards	Neural network analysis (supervised learning) and Gaussian mixture models (unsupervised learning)	To determine whether the shape and size of the snail's shells can distinguish between two <i>Placostylus</i> species, particularly in groups that appear to be hybrids.	Combining geometric morphometrics and machine learning effectively answers biological issues, providing insights into species classification and possible hybridization.	Difficulty classifying intermediate phenotypes, with potential for overfitting and misclassification in both learning methods.
V. M. Tuset and E. Galimany and A. Farrés and E. Marco-Herrero and J. L. Otero-Ferrer and A. Lombarte and M. Ramón	Wavelet functions and Elliptic Fourier descriptors	Addresses the difficulty of accurately defining phenotypic diversity in gastropod shells.	Advanced contour analysis methods allow accurate differentiation of gastropod shell forms.	Cannot clarify the causes of phenotypic variation in the two populations studied.
Fedor Lishchenko and Jones, J. B.	Landmark- and outline-based Geometric Morphometric methods	To address difficulties in differentiating between stocks of marine organisms to prevent misidentification that could affect conservation and management.	Shape analysis improves taxonomic classification precision and offers close distinction between related species or organisms.	Landmark-based methods can be sensitive to landmark placement.
M. Tsutsumi and N. Saito and D. Koyabu and C. Furusawa	Morphological regulated variational AutoEncoder (Morpho-VAE)	The need for reliable, landmark-free methods, such as a modified variational autoencoder, to extract and decipher complex shapes from image data.	Employs dimension reduction and feature extraction, making it a user-friendly tool for biology non-experts.	Limited sample size in certain families presented challenges.
Barrera-Hernandez, R. and Barrera-Soto, V. and Martinez-Rodriguez, J. L. and Ríos-Alvarado, A. B. and Ortiz-Rodríguez, F.	Machine learning algorithms	Identifying the sex of abalones is challenging for producers applying specific growth or preservation strategies.	Machine learning algorithms accurately classify abalone sex into three categories: male, female, and immature.	Selected features may not fully capture the complexity of abalone morphology.
Concepcion, R. and Guillermo, M. and Tanner, S. E. and Fonseca, V. and Duarte, B.	EfficientNet-Bo, ResNet101, MobileNetV2, InceptionV3	Addresses the difficulty of accurately tracing bivalve harvesting origins using computer vision and machine learning algorithms to enhance seafood traceability and combat food fraud.	Non-invasive, image-based tools for bivalve traceability provide faster, cheaper, and equally accurate alternatives to traditional chemical analysis methods.	Small sample size (only 30 cockles) limits model reliability.

Table 2.1: Comparison of the methods used in bivalves studies.

883 Recent developments and breakthroughs in machine learning offer promising solu-
884 tions to biological challenges. Research findings indicate that various deep learning
885 techniques — such as convolutional neural networks (CNNs), geometric morpho-
886 metrics, and other machine learning models — are effective in identifying phe-
887 notypes and determining the sex of various aquaculture species, including mol-
888 lusks and abalones. These techniques provide a foundation for developing new,
889 non-invasive methods to differentiate male and female *T. granosa*, potentially ad-
890 dressing the limitations of manual and invasive techniques. Thus, using machine
891 learning to analyze morphological and morphometric features may streamline the
892 process of sex identification.

893 Nevertheless, the use of machine learning and deep learning to determine the sex
894 of *T. granosa* has not been fully explored. It lacks up-to-date and significant
895 related literature on using machine learning and deep learning to identify sex in
896 *T. granosa*, particularly given the species' possible sequential hermaphroditism
897 and lack of obvious external sexual distinctions.

⁸⁹⁸ **Chapter 3**

⁸⁹⁹ **Research Methodology**

⁹⁰⁰ This chapter discusses the materials and methods employed, focusing on the de-
⁹⁰¹ tailed workflow in conducting the study from sample collection, preprocessing,
⁹⁰² model training and evaluation.

⁹⁰³ Dr. Victor Emmanuel Ferriols, the director of the Institute of Aquaculture, over-
⁹⁰⁴ saw the overall workflow by providing baseline characteristics of the samples that
⁹⁰⁵ the researchers could focus on. Additionally, guidance was offered by the re-
⁹⁰⁶ search associates Allena Esther Artera and LC Mae Gasit. Consequently, the
⁹⁰⁷ entire dataset collection process was conducted at the University of the Philip-
⁹⁰⁸ pines Visayas hatchery facility.

⁹⁰⁹ The methodology consisted of nine parts: (1) Sample Collection, (2) Ethical Con-
⁹¹⁰ siderations, (3) Creating *T.granosa* Dataset, (4) Morphological Characteristics
⁹¹¹ Collection (5) Image Acquisition and Pre-processing, (6) Hardware and Software
⁹¹² Configuration, (7) Machine Learning for Morphometric Data, (8) Deep Learning

913 for Morphological Analysis, and (9) Evaluation Metrics

914 **3.1 Sample Collection**

915 The collection of *T. granosa* samples used in this study was part of an ongoing
916 research project by UPV DOST-PCAARRD titled "Establishment of the Center
917 for Mollusc Research and Development: Development of Spawning and Hatchery
918 Techniques for the Blood Cockle (*Anadara granosa*) for Sustainable Aquaculture."

919 A total of 271 samples were provided for this study to classify the sex of *T.*
920 *granosa*. The samples, ranging in size from 34 to 61 mm, were sourced from the
921 coastal area of Zarraga, Iloilo, and fish markets in Ivisan, Capiz, Philippines (see
922 *Figure 3.1*, and *Figure 3.2*).

923 The research and experimentation were conducted at the University of the Philip-
924 pines Visayas hatchery facility in Miagao, Iloilo, where the samples were main-
925 tained in 200 L fiberglass-reinforced plastic (FRP) tanks containing filtered sea-
926 water with 35 ppt salinity (Miranda & Ferriols, 2023).

927 As part of the data collection process, the researchers utilized induced spaw-
928 ning and dissection to classify the sex of the samples. Induced spawning through
929 temperature fluctuations was the most natural and least invasive method for bi-
930 valves compared to other approaches (Aji, 2011). However, since not all samples
931 exhibited gamete release, the researchers also performed dissections, assisted by
932 hatchery staff, to expedite data collection. The sex of the dissected samples was
933 identified based on the coloration of gonad tissue, which varies according to sex
934 and maturity stage. Females exhibited orange-red to pale orange gonads, while

935 males displayed white to grayish-white gonads (May et al., 2021).

936 The methods used for data collection were considered non-invasive, given that
937 *T. granosa* are oxygen regulators well adapted to tidal exposure and hypoxia
938 (Davenport & Wong, 1986).



Figure 3.1: Female *T. granosa* shells.



Figure 3.2: Male *T. granosa* shells.

939 3.2 Ethical Considerations

940 The ongoing research project titled "Establishment of the Center for Mollusc Re-
941 search and Development: Development of Spawning and Hatchery Techniques for
942 the Blood Cockle (*Anadara granosa*) for Sustainable Aquaculture"—from which
943 the samples used in this study were obtained—was reviewed and approved by the
944 Institutional Animal Care and Use Committee (IACUC) of the University of the
945 Philippines Visayas.

946 3.3 Creating *T. granosa* Dataset

947 The experiment began with the collection of preliminary observations from 100 *T.*
 948 *granosa* samples. For the actual experimentation, the researchers collected the full
 949 dataset in batches until a total sample size of 271 *T. granosa* was reached. Lin-
 950 ear measurements—including width, height, length, rib count, hinge line length,
 951 and the distance between the umbos—were recorded and organized into a CSV
 952 file. This dataset served as the foundation for training and testing machine learn-
 953 ing models, as well as for establishing a baseline for the Convolutional Neural
 954 Networks.

955 Images of each sample were captured and saved in JPG format using a standard-
 956 ized file naming convention that included the sample’s sex, the shell’s orientation
 957 or view, and its corresponding number out of the 271 total samples. File names
 958 for female *T. granosa* samples began with “0”, while those for male samples began
 959 with “1”. Each file name also included one of the six captured views: (1) dorsal,
 960 (2) ventral, (3) anterior, (4) posterior, (5) left lateral, and (6) right lateral (*refer to*
 961 *Figure 3.3*), followed by a unique sample number. For example, “010001” denoted
 962 the first female sample taken from the dorsal view, while “110001” represented the
 963 first male sample from the same view. This naming convention was implemented
 964 to prevent data leakage and ensure accurate labeling of images according to their
 965 respective samples.

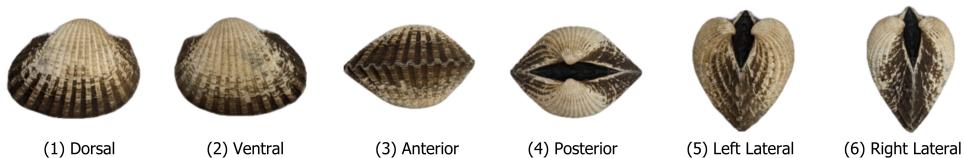


Figure 3.3: Different views of the *T. granosa* shell captured.

966 3.4 Morphometric Data Collection

967 Morphology refers to biological form and is one of the most visually recognizable
968 phenotypes across all organisms (Tsutsumi, Saito, Koyabu, & Furusawa, 2023).
969 In this study, morphological characteristics describe the structural features of
970 *T. granosa*, focusing on measurable attributes such as shape, size, and color.
971 Morphometric characteristics, on the other hand, refer to specific quantifiable
972 features of *T. granosa*, including length, width, height, hinge line length, distance
973 between the umbos, and rib count. As stated by the researchers, quantifying and
974 characterizing these traits is essential for understanding and visualizing variations
975 in *T. granosa* morphology.

976 The researchers measured the height, width, and length of *T. granosa* using a
977 Vernier caliper with a precision of up to 0.01 mm. Refer to Figure 3.4 for the
978 corresponding measurement diagram. Length (A) refers to the distance from the
979 anterior to the posterior of the shell. Width (B) is defined as the widest span
980 across the shell from the left to the right valve. Height (C) measures the distance
981 from the base to the apex of the shell. In addition, the hinge line length (D) near
982 the hinge and the distance between the umbos (E) were recorded.

983 Reyment and Kennedy (1998) emphasized that including rib count (F) as supple-
984 mentary information can enhance identification accuracy. Following this insight,
985 the researchers also recorded the rib count for both male and female *T. granosa*,
986 adjusting the values by calculating ratios to account for natural size variation
987 among specimens.

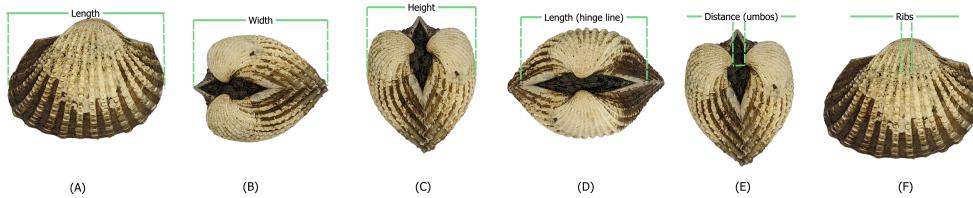


Figure 3.4: Linear measurements that were gathered from the shell of *T. granosa*.

988 3.5 Image Acquisition and Data Gathering

989 This study comprised 144 male and 127 female *T. granosa* samples, resulting
 990 in a total of 1,626 images captured from various angles. To ensure consistency
 991 during image acquisition, the researchers constructed a box-like structure with
 992 a white background to control the imaging environment (see *Figure 3.5*). This
 993 setup allowed for uniform image captures by fixing the camera at a consistent
 994 angle directly above the *T. granosa*. A ring light was positioned in front of the
 995 box to enhance image quality, eliminate shadows, and ensure clarity of the samples
 996 throughout the image acquisition process.

997 The images were captured using a Google Pixel 3 XL smartphone, which features
 998 a resolution of 2960×1440 pixels and a 12.2 MP camera (4032×3024 pixels).
 999 Additional camera specifications include an f/1.8 aperture, 28mm wide lens, $\frac{1}{2.55}$ "
 1000 sensor size, $1.4\mu\text{m}$ pixel size, dual-pixel phase detection autofocus (PDAF), and
 1001 optical image stabilization (OIS) (Concepcion et al., 2023).



Figure 3.5: Image acquisition setup for *T. granosa* samples.

1002 3.6 Hardware and Software Configuration

1003 This section of the paper discusses the software, programming languages, and tools
1004 used for sex identification. Data collection, preprocessing, and model training
1005 were conducted on a Windows 11 operating system using an ACER Aspire 3
1006 general-purpose unit (GPU) equipped with an AMD Ryzen 3 7320U CPU with
1007 Radeon Graphics (8 cores) @ 2.395 GHz and 8 GB of RAM. Google Colaboratory
1008 was utilized for collaborative preprocessing, computer vision tasks, and model
1009 training. Image preprocessing was performed using computer vision techniques in
1010 Python, while machine learning and deep learning models were developed using
1011 Python libraries, including Keras. The results of the gathered measurements were
1012 stored and managed using spreadsheet software. GitHub was employed for version
1013 control, documentation, and activity tracking throughout the study.

3.7 Machine Learning on Morphometric Data

This section of the paper discusses the machine learning operations that served as a baseline prior to implementing more complex deep learning methods for image classification. The study utilized collected variables including linear measurements—length, width, height, hinge line length, distance between the umbos, and rib count—along with derived features used as predictors. These included the length-to-width ratio, length-to-height ratio, width-to-height ratio, umbos distance-to-length ratio, hinge line length-to-length ratio, umbos distance-to-height ratio, and rib density. The samples were classified by sex, with females labeled as 0 and males as 1, which served as the response variable.

1024 **3.7.1 Data Preprocessing**

1025 The preprocessing of the dataset involved several essential steps, carried out using
1026 Python in Google Colaboratory, in preparation for machine learning analysis (*see*
1027 *Figure 3.6*).

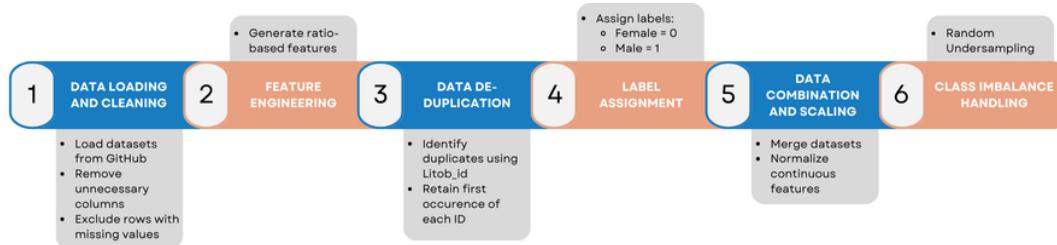


Figure 3.6: Data preprocessing in machine learning pipeline.

1028 **Data Loading and Cleaning**

1029 The process began by loading two separate datasets for male and female *T. granosa*
1030 directly from GitHub using `pd.read_csv()`. Unnecessary columns were removed,
1031 and rows containing missing values were excluded using the `dropna()` function to
1032 ensure data completeness and reliability.

1033 **Feature Engineering**

1034 Additional ratio-based features were generated to augment the existing measurements. These included the length-to-width ratio, length-to-height ratio, width-to-height ratio, hinge line length-to-length ratio, umbos distance-to-length ratio, umbos distance-to-height ratio, and rib density. These derived features aimed to emphasize shape characteristics independent of size, improving the models' ability 1038 to distinguish morphological differences between sexes.

1040 ***Data De-duplication***

1041 To avoid redundancy and ensure each specimen was uniquely represented, the
1042 last three digits of each `Litob_id` were used to identify duplicates. Only the first
1043 occurrence of each unique ID was retained, reducing potential bias caused by
1044 repeated entries.

1045 ***Label Assignment***

1046 A new column labeled `Label` was added to both datasets. Female specimens were
1047 assigned a label of 0, and male specimens a label of 1. This column served as the
1048 target variable for classification.

1049 ***Data Combination and Scaling***

1050 After cleaning and feature engineering, the male and female datasets were merged
1051 into a single DataFrame. The `Litob_id` column was removed post de-duplication.
1052 All continuous numeric features were normalized using `MinMaxScaler` to scale
1053 values to the range [0, 1].

1054 Rib count was excluded from normalization because it is a discrete feature with
1055 biologically meaningful bounds. According to best practices in machine learning,
1056 normalizing discrete or categorical features can distort their meaning and is often
1057 unnecessary (Jaiswal, 2024). In this study, rib count was treated as a categorical
1058 attribute due to its biological significance and finite, non-continuous nature.

1059 ***Class Imbalance Handling***

1060 After normalization, class imbalance was addressed by applying Random Under-
1061 sampling to the male dataset. This technique randomly reduced the number of

1062 male samples to match the number of female samples (127 each), ensuring equal
1063 class representation. By using this approach, model bias was minimized, and the
1064 classification performance became more reliable across both classes.

1065 3.7.2 Machine Learning Models Training

1066 *Model Selection and Hyperparameter Tuning*

1067 To establish a baseline for classification, various models were evaluated: Logis-
1068 tic Regression, K-Nearest Neighbors, Support Vector Machine, Random Forest,
1069 AdaBoost, Extra Trees, and Gradient Boosting. Hyperparameter tuning was con-
1070 ducted using `GridSearchCV`, which systematically identified the optimal settings
1071 for each model to enhance accuracy and performance.

1072 *Cross-Validation*

1073 A five-fold cross-validation approach was implemented (*refer to Figure 3.7*). The
1074 dataset was divided into five subsets, with four used for training and one for
1075 validation. This process was repeated five times, with each fold serving as the
1076 validation set once. This method ensured that model evaluation was robust and
1077 generalizable, minimizing the bias that may result from a single train-test split.

1078 (GeeksforGeeks, 2024)

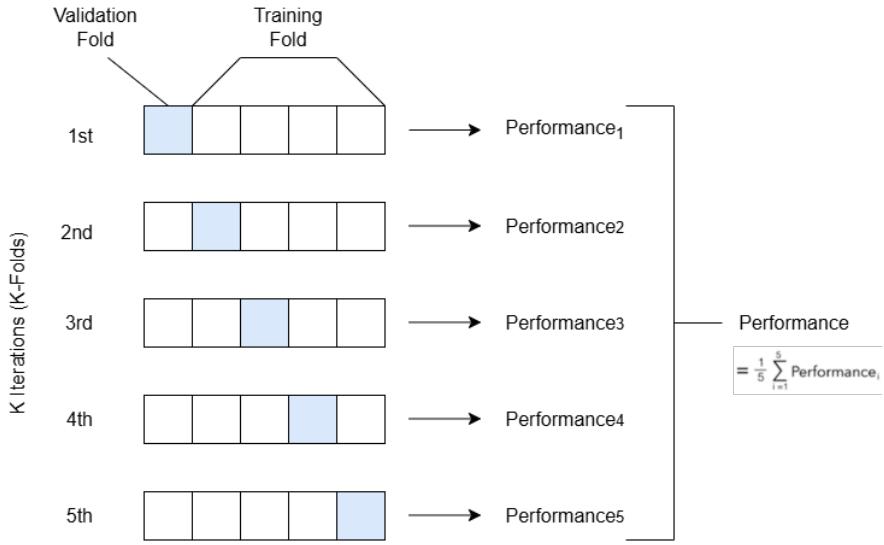


Figure 3.7: Diagram of k-fold cross-validation with $k = 5$.

¹⁰⁷⁹ 3.8 Deep Learning for Morphological Analysis

¹⁰⁸⁰ This section outlines the application of deep learning techniques in analyzing the
¹⁰⁸¹ morphological characteristics of *Tegillarca granosa* to identify their sex based on
¹⁰⁸² shell images. A Convolutional Neural Network (CNN) architecture was imple-
¹⁰⁸³ mented and trained on preprocessed images using stratified cross-validation.

¹⁰⁸⁴ 3.8.1 Image Preprocessing

¹⁰⁸⁵ This subsection details the image processing techniques applied to raw shell images
¹⁰⁸⁶ of *T. granosa* using computer vision methods before training the deep learning
¹⁰⁸⁷ model. The image preprocessing techniques include standardizing input dimen-
¹⁰⁸⁸ sions and removing shadows, background, and noise. Each image underwent data
¹⁰⁸⁹ augmentation to enhance feature visibility for effective learning. Image prepro-
¹⁰⁹⁰ cessing ensures consistent and high-quality input data for model training.

1091 *Adjusting Dimensions*

1092 All images were resized to a consistent dimension of 256x256 pixels to ensure
1093 uniformity throughout the dataset. This standardization is essential for Convo-
1094 lutional Neural Networks (CNNs), as a consistent input dimension is required.
1095 While resizing, the aspect ratio was maintained to prevent distortion of the mor-
1096 phological features, and padding was added to retain the original format.

1097 *Background Removal*

1098 Background removal was performed to maintain a consistent white background
1099 throughout the dataset. The tool `rembg` was used to efficiently remove the original
1100 background, retaining the foreground from the raw images. This method resulted
1101 in clear images with a white background, enhancing focus on the morphological
1102 features and defining the shell boundaries.

1103 *Shadow Removal*

1104 To minimize noise caused by shadows around the shell, HSV thresholding, con-
1105 tours, and morphological thresholds were applied to isolate and remove shadowed
1106 regions. This approach preserved the natural color of the blood cockles and elim-
1107 inated shadows and noise from the surrounding area (*see Figures 3.8 and 3.9*).

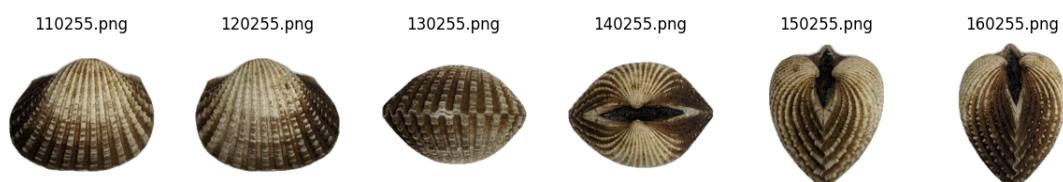


Figure 3.8: Shadows removed from male samples at different angles.



Figure 3.9: Shadows removed from female samples at different angles.

1108 3.8.2 Convolutional Neural Network

1109 Convolutional Neural Networks are the deep learning tool used in image classifica-
1110 tion, specifically binary classification. CNNs leverage their ability to share weights
1111 and use pooling techniques, reducing the number of parameters (Cui, Pan, Chen,
1112 & Zou, 2020). The proposed CNN architecture for sex identification of blood
1113 cockles employs 5 layers, designed to extract features from the input image with
1114 dimensions. The layers consist of three convolution layers, a pooling layer, a flat-
1115 ten layer, dropout, and two dense layers. The CNN framework used in this study
1116 was adapted from an open source GitHub implementation by Christian Versloot,
1117 which focused on K-fold Cross Validation using TensorFlow and Keras, which was
1118 customized to align with the objectives of this study. The overall framework is
1119 illustrated in Figure 3.10.

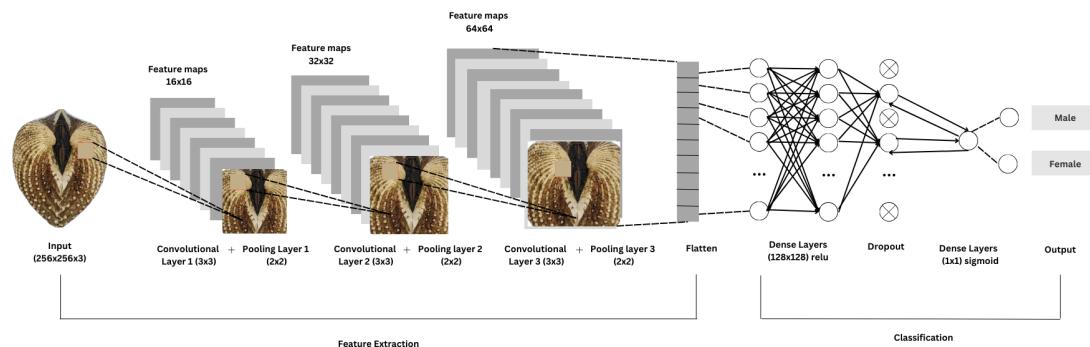


Figure 3.10: Architecture of convolutional neural network (CNN).

1120 ***Convolution Layer***

1121 The convolution layers of CNN extract the features from the input image through
1122 the convolution operation. This study uses three convolution layers with a 3x3
1123 kernel size and filter sizes of 16, 32, and 64 (*refer to Table 3.1*). The first layer
1124 extracts the low-level features, such as edges, lines, and corners, while the deeper
1125 layers iteratively extract more complex information from these low-level features.
1126 The ReLU activation function is used as the baseline for this model, and experi-
1127 ments are conducted with different activation functions, such as ELU and PReLU,
1128 to evaluate their impact on learning complex patterns within the data.

1129 ***Pooling Layer***

1130 A pooling layer was added after the convolution layer to enhance calculation speed
1131 and prevent overfitting (Cui et al., 2020). In this study, max pooling was applied
1132 with a (2,2) kernel size.

1133 ***Fully Connected and Dropout***

1134 Fully connected layers follow after the convolution and pooling layers. Each neu-
1135 ron connects to all neurons of the previous layer. The output values from the
1136 fully connected layers are sent to an output layer. It was classified using different
1137 sigmoid functions appropriate for binary classification.

1138 A large number of parameters in the training process can lead to overfitting. It
1139 occurs when the model learns the training data too well, including its noise and
1140 irrelevant details. This results in poor performance on unseen data. To mitigate
1141 the overfitting, the dropout layer was employed. Dropout works by temporarily
1142 discarding a portion of the neurons in the network with probability p ($0 < p < 1$).

- ¹¹⁴³ During this process, these neurons do not participate in the forward propagation
¹¹⁴⁴ process of CNN and the backward propagation process (Cui et al., 2020).

Layer	Number of Neurons	Stride	Kernel Size	Activation	Parameters
Rescaling					
Convolution	16	1x1	3x3	ReLU	448
Max Pooling		1x1	2x2		
Convolution	32	1x1	3x3	ReLU	4,640
Max Pooling		1x1	2x2		
Convolution	64	1x1	3x3	ReLU	18,496
Max Pooling		1x1	2x2		
Flatten					
Dense	128			ReLU	7,372,928
Dropout					
Dense	1			Sigmoid	129

Table 3.1: Architecture of the convolutional neural network used.

¹¹⁴⁵ 3.8.3 CNN Training

- ¹¹⁴⁶ The dataset consists of 1626 images, with 127 samples from females and 144 sam-
¹¹⁴⁷ ples from males, individually for each angle. Given the minimal class imbalance,
¹¹⁴⁸ random undersampling was carried out to create a balanced dataset. All images
¹¹⁴⁹ were resized to 256x256 pixels and normalized using a Rescaling layer, ensuring
¹¹⁵⁰ pixel values were within the range [0, 1].

¹¹⁵¹ *Data Splitting*

- ¹¹⁵² Due to the limited dataset size, a traditional train-test split was not adopted.
¹¹⁵³ Instead, a 5-fold stratified cross-validation approach was used to maximize the
¹¹⁵⁴ use of available data while preserving the class distribution within each fold (*refer*
¹¹⁵⁵ *to Figure 3.11*). `StratifiedKFold` was applied to ensure that the distribution of
¹¹⁵⁶ male and female samples remained consistent across all folds, thereby enabling
¹¹⁵⁷ fair and robust model evaluation (GeeksforGeeks, 2020).

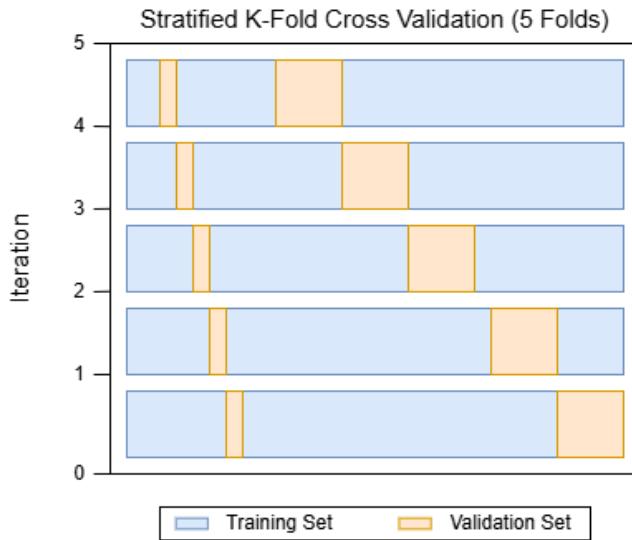


Figure 3.11: Diagram of stratified k-fold cross-validation with $k=5$.

1158 ***Data Augmentation***

1159 Before model training, online data augmentation was applied exclusively to the
1160 training data within each fold, creating new data variations on the fly. The aug-
1161 mentations included random horizontal flips, slight rotations, and zoom trans-
1162 formations to enhance data diversity and improve model generalization (Awan,
1163 2022). All augmentation was strictly applied only to the training subset of each
1164 fold to prevent data leakage and maintain the validity of the results (*Figure 3.12*).

1165 On-the-fly data augmentation (OnDAT) generates augmented data during each
1166 iteration, exposing the model to constantly changing data variations. Augmenting
1167 the original data allows better exploration of the underlying data generation pro-
1168 cess and has the potential to prevent the model from overfitting spurious patterns,
1169 thereby improving performance (Cerqueira, Santos, Baghoussi, & Soares, 2024).

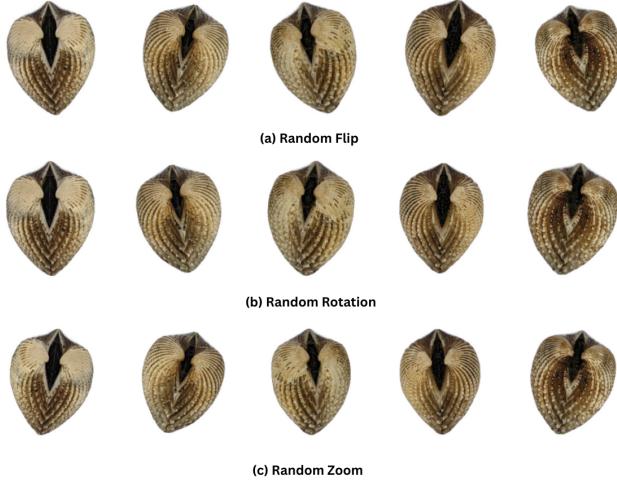


Figure 3.12: On-the-Fly dataset augmentation (OnDAT) techniques.

1170 *Training Procedure*

1171 During the training process, model performance per fold was carefully monitored.
1172 One important thing to observe is the consistency in the performance, whether
1173 the model is still learning or is at high risk of overfitting. Early stopping was ap-
1174 plied to ensure the stable performance of the model across folds. This technique
1175 allows for monitoring the training of the neural network, stopping when the per-
1176 formance metrics, in this case, validation loss, cease to improve. Furthermore, to
1177 enhance the learning process, `ReduceLROnPlateau` was applied, which decreased
1178 the learning rate if there was no improvement in the model for a specified number
1179 of epochs (Team, n.d.).

1180 The model was trained using the Adam optimization algorithm, with an initial
1181 learning rate of 0.001. Binary cross-entropy, commonly known as the log loss,
1182 was employed as the loss function due to its effectiveness in binary classification
1183 tasks. To reduce the risk of overfitting, a dropout rate of 0.5 was applied, ran-
1184 domly deactivating half of the neurons during the training process to improve

1185 generalization.

1186 3.9 Evaluation Metrics

1187 Evaluating the performance of a binary classification model is essential, and se-
1188 lecting appropriate metrics depends on the specific requirements of the user. The
1189 performance of both supervised machine learning and deep learning models will
1190 be measured using several key metrics, including accuracy, precision, recall, F1
1191 Score, and the AUC-ROC Score.

1192 Accuracy (ACC) is the ratio of the overall correctly predicted samples to the
1193 total number of examples in the evaluation dataset (Cui et al., 2020). It measures
1194 the overall correctness of the model in predicting both male and female blood
1195 cockles. This metric provides insight into how well the model performs across all
1196 classifications. The formula for accuracy is:

$$\text{ACC} = \frac{\text{Correctly classified samples}}{\text{All samples}} = \frac{TP + TN}{TP + FP + TN + FN} \quad (3.1)$$

1197 where:

1198 TP or true positive is the number of male samples that were correctly iden-
1199 tified as male *T. granosa*,

1200 TN or true negative is the number of female samples that were correctly iden-
1201 tified as female *T. granosa*,

1202 FP or false positive is the number of female samples that were incorrectly
1203 identified as male *T. granosa*, and

₁₂₀₄ FN or false negative is the number of male samples that were incorrectly
₁₂₀₅ identified as female *T. granosa*.

₁₂₀₆ Precision (PREC) is the ratio of correctly predicted positive samples to all samples
₁₂₀₇ assigned to the positive class (Cui et al., 2020). This metric helps in evaluating
₁₂₀₈ the fairness of the model and prevents the misclassification of blood cockles as it
₁₂₀₉ identifies potential inaccuracies or biases. The formula for precision is:

$$\text{PREC} = \frac{\text{True positive samples}}{\text{Samples assigned to positive class}} = \frac{TP}{TP + FP} \quad (3.2)$$

₁₂₁₀ Recall (REC), also known as sensitivity or the true positive rate (TPR), is the
₁₂₁₁ ratio of correctly predicted positive cases to all the actual positive samples (Cui
₁₂₁₂ et al., 2020). It represents the ability of the model to correctly identify positive
₁₂₁₃ male and female samples. The formula for recall is:

$$\text{REC} = \frac{\text{True positive samples}}{\text{Samples classified positive}} = \frac{TP}{TP + FN} \quad (3.3)$$

₁₂₁₄ The F1 Score is the harmonic mean of precision and recall, which penalizes extreme
₁₂₁₅ values of either of the two metrics (Cui et al., 2020). It is particularly useful when
₁₂₁₆ the class distribution is imbalanced. The formula for the F1 Score is:

$$\text{F1} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = \frac{2 \times TP}{2 \times TP + FP + FN} \quad (3.4)$$

₁₂₁₇ The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) is a
₁₂₁₈ performance measurement for classification problems. The ROC curve is a plot of

₁₂₁₉ the true positive rate (recall) against the false positive rate (1 - specificity), and
₁₂₂₀ the AUC score quantifies the overall ability of the model to discriminate between
₁₂₂₁ positive and negative classes. A higher AUC indicates better model performance.
₁₂₂₂ (Nahm, 2022)

₁₂₂₃ **Chapter 4**

₁₂₂₄ **Results and Discussions**

₁₂₂₅ This chapter presents the results from the machine learning and deep learning
₁₂₂₆ analyses conducted on the preprocessed dataset. It includes an evaluation of
₁₂₂₇ various machine learning classifiers and the application of deep learning models
₁₂₂₈ for image-based classification. The primary focus is on identifying key predictors
₁₂₂₉ and assessing classification performance for sex identification in *T. granosa*.

₁₂₃₀ **4.1 Machine Learning Analysis**

₁₂₃₁ This chapter outlines the results of preprocessing, training of machine learning
₁₂₃₂ models, and feature importance analysis, all conducted in Google Colab using
₁₂₃₃ Python. The dataset was preprocessed in Colab, and the training and evaluation
₁₂₃₄ of various classifiers were performed entirely within this environment. This part
₁₂₃₅ of the paper includes five subsections: data exploration, statistical analysis, fea-
₁₂₃₆ ture importance analysis, performance evaluation, and visualizations for machine

₁₂₃₇ learning.

₁₂₃₈ 4.1.1 Data Exploration

₁₂₃₉ Exploratory data analysis was performed to characterize the dataset using visu-
₁₂₄₀ alizations to understand the patterns and correlations within the data. A corre-
₁₂₄₁ lation heatmap was created to assess the relationship between the predictors and
₁₂₄₂ the target variable.

₁₂₄₃ The heatmap (*see Figure 4.1*) revealed three features most correlated with the
₁₂₄₄ sex of *T. granosa*: the width-height ratio ($r = 0.18$), the umbos distance-length
₁₂₄₅ ratio ($r = 0.12$), and the distance between the umbos ($r = 0.12$). Each of these
₁₂₄₆ features demonstrated a weak positive relationship with the target variable.

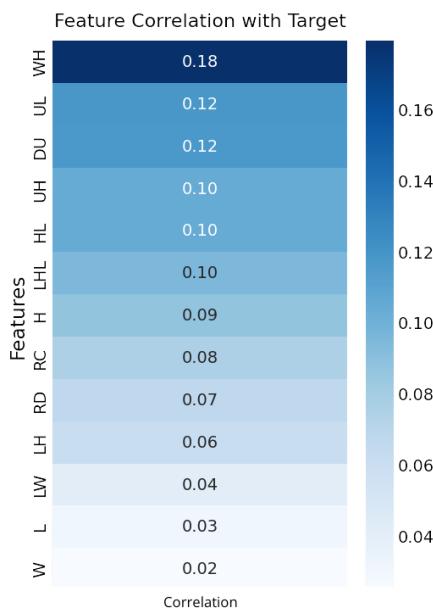


Figure 4.1: Heatmap of morphometric correlations with *T. granosa* sex.

₁₂₄₇ **4.1.2 Statistical Analysis**

₁₂₄₈ As part of the exploratory data analysis, statistical testing confirmed that the
₁₂₄₉ dataset did not follow a normal distribution (*see Table 4.1*). Consequently, the
₁₂₅₀ Mann-Whitney U test was applied with a significance level of $\alpha = 0.05$ to compare
₁₂₅₁ male and female samples. Out of thirteen features, five showed statistically sig-
₁₂₅₂ nificant differences. These included: width-height ratio ($p = 0.003$), length-width
₁₂₅₃ ratio ($p = 0.011$), umbos distance-length ratio ($p = 0.019$), distance between
₁₂₅₄ umbos ($p = 0.025$), and umbos distance-height ratio ($p = 0.036$).

₁₂₅₅ It is important to note that statistical significance does not imply predictive im-
₁₂₅₆ portance. Therefore, further analysis, such as feature importance evaluation, was
₁₂₅₇ performed to identify the most informative predictors for classification.

Variable	p-value
WH_ratio	0.003
LW_ratio	0.011
UL_ratio	0.019
Distance Umbos	0.025
UH_ratio	0.036
HL_ratio	0.079
Length (Hinge Line)	0.120
Height	0.124
Rib Density	0.181
Rib count	0.251
Length	0.334
LH_ratio	0.490
Width	0.753

Table 4.1: Mann-Whitney U test results for sex-based feature comparison.

1258 4.1.3 Feature Importance Analysis

1259 Feature importance was assessed using the Kruskal-Wallis test, a non-parametric
 1260 method that is suitable for evaluating differences in distributions across groups
 1261 when the data does not follow a normal distribution. This approach was chosen
 1262 because of the non-normality of the dataset and its robustness in handling con-
 1263 tinuous and ordinal data without assuming homogeneity of variances. (Ribeiro,
 1264 2024)

1265 Kruskal-Wallis test analysis showed that the width-to-height ratio (WH ratio)
 1266 had the highest importance score, indicating it is the most statistically significant
 1267 feature for distinguishing the sex of *T. granosa*. Other notable features included
 1268 the length-to-width ratio (LW ratio), umbos distance-to-length ratio (UL ratio),
 1269 distance between the umbos, and umbos distance-to-height ratio (UH ratio), all
 1270 of which contributed significantly to the classification task (*refer to Figure 4.2*).

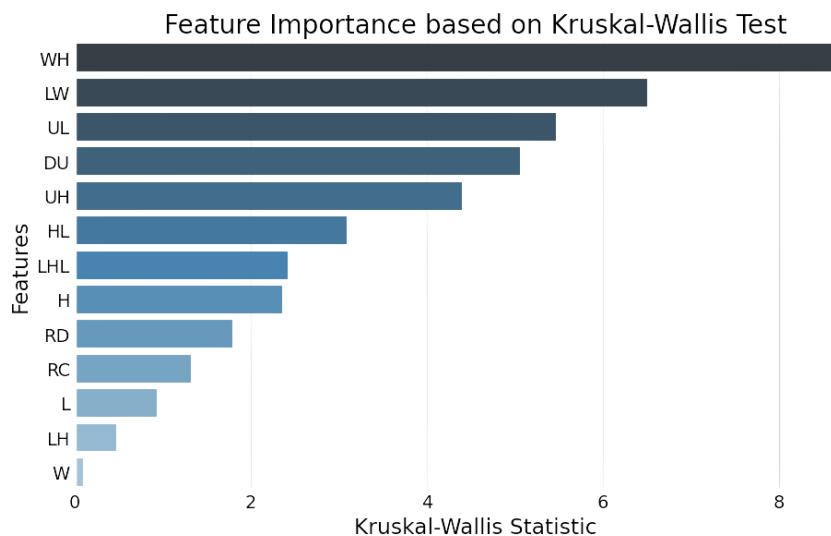


Figure 4.2: Feature importance scores using the Kruskal-Wallis test.

4.1.4 Performance Evaluation

Table 4.2 shows the performance metrics of different machine learning models trained using all 13 features from the dataset. Among the models, Gradient Boosting achieved the highest accuracy of 61.03%, along with strong precision, recall, and F1 Score values. AdaBoost also performed competitively, with an accuracy of 60.63%. These results highlight the effectiveness of ensemble methods such as Gradient Boosting and AdaBoost when utilizing the full feature set, likely because of their capability to combine multiple weak learners into a more robust predictive model (Hussain & Zaidi, 2024).

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Support Vector Machine	58.62	58.62	58.62	58.44
Logistic Regression	57.83	57.83	57.83	57.61
K-Nearest Neighbors	51.18	51.31	51.18	50.77
Extra Trees	59.07	59.54	59.07	58.45
Random Forest	59.85	59.99	59.85	59.80
Gradient Boosting	61.03	61.32	61.03	60.81
AdaBoost	60.63	60.98	60.63	60.39

Table 4.2: Performance metrics for models with all 13 features.

Table 4.3 presents the performance of the same models using only the top five features identified through Kruskal-Wallis feature importance analysis. The selected features are the distance between the umbos, length-to-width ratio, width-to-height ratio, umbos distance-to-height ratio, and umbos distance-to-length ratio.

Interestingly, the overall performance of the models improved when using only the top 5 features compared to using all 13. K-Nearest Neighbors (KNN) achieved the best results with an accuracy of 64.16%, precision of 64.97%, recall of 64.16%, and an F1 Score of 63.75%. Gradient Boosting followed closely behind. These findings suggest that reducing the feature set to the most relevant variables helped

1289 simplify the models, improved generalization, and enhanced predictive performance—particularly for KNN, which showed a notable improvement over its earlier results with the full feature set.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Support Vector Machine	63.77	64.47	63.77	63.42
Logistic Regression	63.75	63.87	63.75	63.70
K-Nearest Neighbors	64.16	64.97	64.16	63.75
Extra Trees	61.04	61.68	61.04	60.67
Random Forest	61.01	61.12	61.01	60.91
Gradient Boosting	64.15	64.24	64.15	64.01
AdaBoost	61.02	61.26	61.02	60.82

Table 4.3: Performance metrics for models with 5 features.

1292 4.1.5 Visualizations for Machine Learning

1293 Figure 4.3 is a confusion matrix that summarizes the performance of the K-Nearest
1294 Neighbors model in classifying *T. granosa* based on their sex, where 0 represents
1295 female samples and 1 represents male samples. From the matrix, it can be ob-
1296 served that out of all the actual female samples (true label 0), 91 were correctly
1297 predicted as female (true positive for class 0), while 36 were incorrectly classified
1298 as male (false negative for class 0). On the other hand, out of all the actual male
1299 samples (true label 1), 72 were correctly predicted as male (true positive for class
1300 1), while 55 were incorrectly classified as female (false negative for class 1).

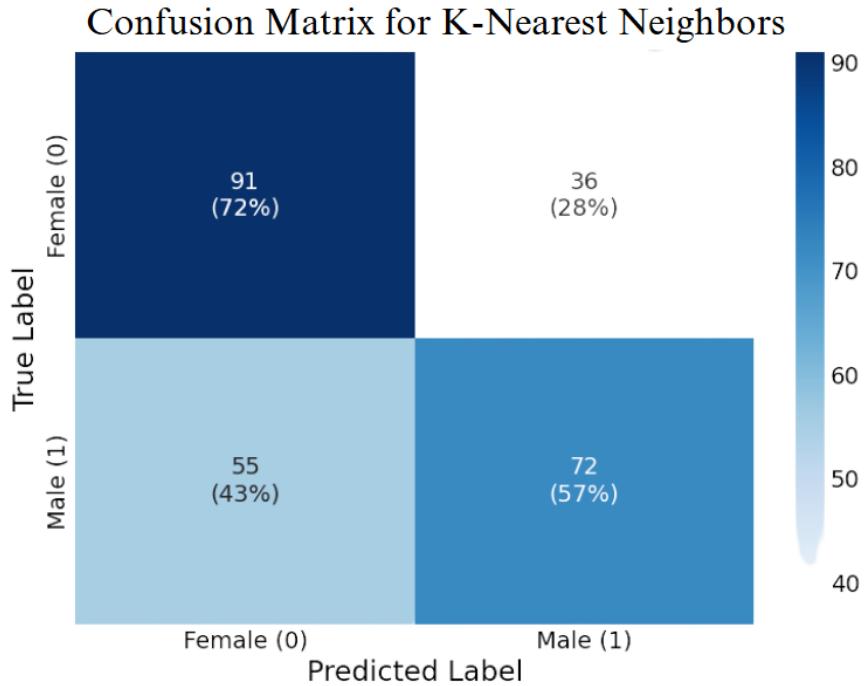


Figure 4.3: KNN confusion matrix for *T. granosa* sex classification.

1301 Figure 4.4 displays the average ROC curve, showing KNN's ability to distinguish
 1302 between positive and negative cases. The ROC curve helps assess the trade-off
 1303 between sensitivity (true positive rate) and specificity (1 - false positive rate).
 1304 The Area Under the Curve (AUC) value, which ranges from 0.5 (random chance)
 1305 to 1 (perfect discrimination), is used to evaluate the model's overall performance.
 1306 In this case, KNN achieved an average AUC of 0.7004, indicating that it performs
 1307 better than random guessing and has reasonable predictive ability.

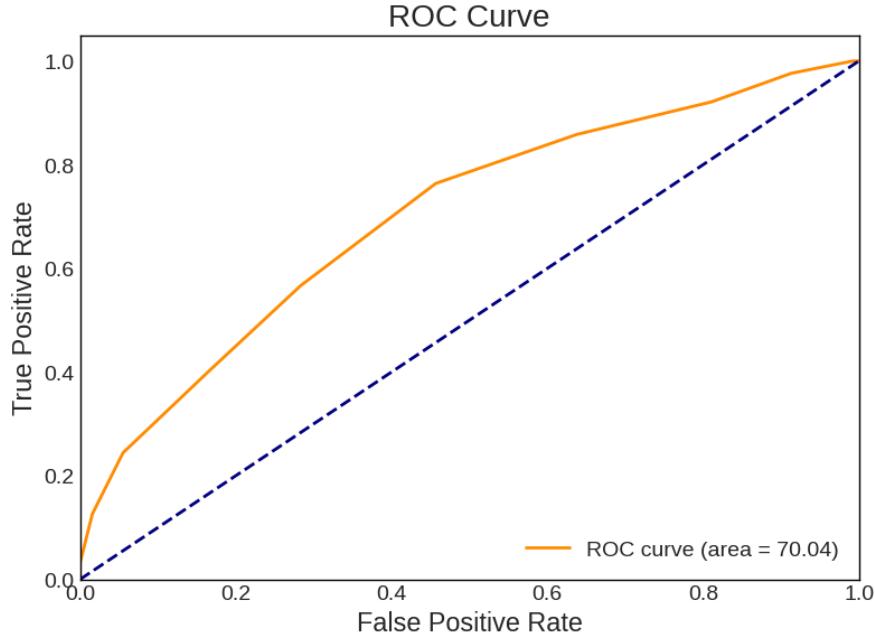


Figure 4.4: ROC curve with AUC score for KNN.

1308 4.2 Deep Learning Analysis

1309 This section presents the performance of the Convolutional Neural Network (CNN)
1310 model in classifying the sex of *T. granosa* based on shell morphology. The analysis
1311 evaluates the model's ability to distinguish between male and female shell images
1312 using various evaluation metrics. This part of the paper includes six subsections:
1313 baseline model, comparison of individual and combined angles, training result and
1314 hyperparameter tuning, proposed model, learning rates and training behavior per
1315 fold, and visualizations for deep learning.

1316 The machine learning analysis (*see Figure 4.3*) revealed that five of the original
1317 features produced significant results. The K-Nearest Neighbors (KNN) model
1318 achieved an accuracy of 64.16%, precision of 64.97%, recall of 64.16%, and an F1
1319 Score of 63.75%. This section compares the model's performance across differ-

₁₃₂₀ ent angles based on the results of the machine learning and feature importance
₁₃₂₁ analysis.

₁₃₂₂ 4.2.1 Baseline Model

₁₃₂₃ This section presents the baseline model with a batch size of 16 and 20 epochs,
₁₃₂₄ which will serve as the starting point for comparison and provide a guideline for
₁₃₂₅ hyperparameter tuning. The focus will be on one of the angles, specifically the
₁₃₂₆ Left Lateral view, since the feature importance analysis using the Kruskal-Wallis
₁₃₂₇ Test indicated that the width-to-height ratio had the highest importance score,
₁₃₂₈ which is most visible from the Left Lateral view.

₁₃₂₉ The unbalanced dataset, which consisted of 144 male samples and 127 female
₁₃₃₀ samples, achieved an accuracy of 65.27%, precision of 71.82%, recall of 58.99%,
₁₃₃₁ an F1 Score of 63.99%, an AUC score of 73.08%, and a loss of 0.6122. However, to
₁₃₃₂ address the class imbalance and enhance model performance, random undersam-
₁₃₃₃ pling was performed. This approach resulted in improved performance metrics for
₁₃₃₄ the balanced dataset, with an accuracy of 67.34%, precision of 69.43%, a recall
₁₃₃₅ of 64.06%, an F1 Score of 65.60%, an AUC score of 74.31%, and a lower loss of
₁₃₃₆ 0.5981.

Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	AUC score (%)	Loss (%)
Unbalanced	65.27	71.82	58.99	63.99	73.08	0.6122
Balanced	67.34	69.43	64.06	65.60	74.31	0.5981

Table 4.4: Performance metrics for unbalanced vs. balanced datasets (Batch Size: 16, Epochs: 20).

4.2.2 Comparison of Individual and Combined Angles

Using the same batch size and number of epochs, performance was compared across all individual angles and the combination of the two highest-performing angles based on accuracy, using a balanced dataset. For the combined analysis, samples from the two selected angles were placed side by side, and a new dataset folder was created for male and female samples.

Table 4.5 presents the performance metrics for each individual angle and the combination of the two highest-performing angles in terms of accuracy. The Left Lateral view achieved the highest accuracy (67.34%) and precision (69.43%), while the Dorsal view obtained the highest recall (77.88%) and F1 Score (69.96%). Meanwhile, the Ventral view recorded the highest AUC score (74.87%), indicating its strong ability to distinguish between classes. Combining the Ventral and Left Lateral views resulted in an overall accuracy of 62.60%, suggesting that while combined images may provide complementary information, individual angle views still outperformed the combined views under the current experimental setup.

Angle	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	AUC score (%)	Loss (%)
Dorsal	66.54	63.76	77.88	69.96	73.09	0.6152
Ventral	67.30	69.33	66.18	66.53	74.87	0.6159
Anterior	51.57	31.11	6.31	10.02	65.87	0.6825
Posterior	61.43	63.48	51.17	54.25	70.12	0.6257
Left Lateral	67.34	69.43	64.06	65.60	74.31	0.5981
Right Lateral	65.37	67.18	59.82	62.99	71.02	0.6115
Ventral + Left Lateral	62.60	67.02	57.85	58.57	70.37	0.6433

Table 4.5: Performance metrics for individual and combined angles (Batch Size: 16, Epochs: 20).

1352 4.2.3 Training Result and Hyperparameter Tuning

1353 The Left Lateral angle view was selected for further optimization. Several ex-
1354 periments were conducted by tuning hyperparameters such as batch size, number
1355 of epochs, and activation functions. Each adjustment was compared against the
1356 baseline model to enhance performance and develop a robust CNN for sex classi-
1357 fication of *T. granosa*.

1358 The Left Lateral angle was chosen because it achieved the highest accuracy and
1359 precision among all individual views, and because the Kruskal-Wallis feature im-
1360 portance analysis indicated that the width-to-height ratio, a feature most visible
1361 from the lateral perspective, was the most significant morphological trait for clas-
1362 sification. Therefore, focusing on this view was expected to maximize the model's
1363 learning capacity and improve classification performance.

1364 A. Batch Size and Number of Epochs

1365 Table 4.6 shows the results indicating that a batch size of 32 with 50 epochs
1366 achieved the best overall performance, with an accuracy of 71.68%, a precision of
1367 72.52%, a recall of 69.29%, an F1 Score of 69.12%, and AUC score of 77.34%.

1368 In contrast, increasing the batch size to 64 resulted in lower recall and F1 Scores,
1369 suggesting that smaller batch Sizes (16 or 32) are more effective for this dataset.
1370 A moderate batch size of 32 allowed the model to generalize better and maintain
1371 stable learning, while too large batch sizes may have led to underfitting.

Epoch	Batch Size	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	AUC Score (%)	Loss
20	16	67.34	69.43	64.06	65.60	74.31	0.5981
	32	68.13	72.25	58.95	62.34	74.76	0.6041
	64	56.71	65.96	36.83	41.46	71.28	0.6692
30	16	67.73	70.17	64.06	65.72	75.76	0.5900
	32	71.28	73.17	66.89	68.27	76.76	0.5832
	64	57.95	61.94	48.12	52.66	71.22	0.6241
32	16	67.73	70.17	64.06	65.72	75.76	0.5900
	32	71.68	72.52	69.29	69.12	77.34	0.5824
	64	61.10	62.68	56.12	56.83	73.46	0.6086

Table 4.6: Effect of batch size and epoch values on CNN model performance.

1372 B. Activation Functions

1373 Table 4.7 shows the performance of different activation functions applied to the
 1374 CNN model trained with a batch size of 32 and 50 epochs. Based on the results,
 1375 the ReLU activation function achieved the best overall performance, with an ac-
 1376 curacy of 71.68%, precision of 72.52%, recall of 69.29%, F1 Score of 69.12%, and
 1377 AUC score of 77.34%, along with the lowest loss at 0.5824. This suggests that
 1378 ReLU remains an effective activation function for the classification of *T. granosa*,
 1379 outperforming both ELU and PReLU in this setup.

Activation Functions	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	AUC score (%)	Loss (%)
ReLU	71.68	72.52	69.29	69.12	77.34	0.5824
ELU	53.14	32.91	53.08	39.95	58.23	0.6796
PreLU	62.64	66.59	50.43	56.96	72.33	0.6162

Table 4.7: Performance metrics for different activation functions (Batch Size: 32, Epochs: 50).

1380 4.2.4 Proposed Model

1381 This section presents the performance evaluation of the proposed Convolutional
 1382 Neural Network (CNN) model, trained with a batch size of 32, 50 epochs, and us-
 1383 ing the ReLU activation function. The model's effectiveness was assessed through
 1384 5-fold cross-validation to ensure robustness and generalizability across different

1385 data partitions.

1386 The proposed model consistently achieved high performance in Folds 1, 3, and
1387 5, with accuracies above 76% and strong recall and AUC scores, demonstrating
1388 its potential for reliable sex identification of *T. granosa*. The slight variation
1389 in performance across folds may be attributed to differences in data distribution,
1390 emphasizing the importance of further data augmentation and balancing for future
1391 work.

Fold no.	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	AUC score (%)	Loss (%)
Fold 1	76.47	70.59	92.31	80.00	73.08	0.5975
Fold 2	62.75	70.59	46.15	55.81	71.85	0.6202
Fold 3	78.43	75.00	84.00	79.25	84.92	0.5392
Fold 4	62.75	71.43	40.00	51.28	71.08	0.6331
Fold 5	78.00	75.00	84.00	79.25	85.76	0.5219

Table 4.8: Per-fold performance metrics (Batch Size: 32, Epochs: 50, Activation Function: ReLU).

1392 4.2.5 Learning Rates and Training Behavior per Fold

1393 This section presents the learning rate adjustments, early stopping events, and
1394 best epoch selections for each fold during the 5-fold cross-validation of the pro-
1395 posed model. During training, the ReduceLROnPlateau callback was employed
1396 to monitor the validation loss and automatically reduce the learning rate when
1397 performance plateaued. Additionally, EarlyStopping was utilized to halt training
1398 once no further improvement was observed after a set patience, and the model
1399 weights were restored from the end of the best-performing epoch to ensure optimal
1400 performance.

1401 The following table summarizes the epochs where learning rate reductions oc-
1402 curred, the adjusted learning rates, the epochs at which early stopping took place,

¹⁴⁰³ and the best epochs from which model weights were restored for each fold.

Fold no.	Epoch (LR Reduced)	Learning Rate After Reduction	Early Stopping Epoch	Best Epoch (Restored)
Fold 1	20	0.0005000	25	17
	23	0.0002500		
Fold 2	9	0.0005000	19	11
	14	0.0002500		
	17	0.0001250		
Fold 3	15	0.0005000	20	12
	18	0.0002500		
Fold 4	12	0.0005000	32	24
	15	0.0002500		
	27	0.0001250		
	30	0.0000625		
Fold 5	20	0.0005000	25	17
	23	0.0002500		

Table 4.9: Learning rate reductions, early stopping, and best epochs per fold during 5-fold cross-validation.

¹⁴⁰⁴ 4.2.6 Visualizations for Deep Learning

¹⁴⁰⁵ Figure 4.5 shows the performance of the model in the training and validation in
¹⁴⁰⁶ terms of accuracy across five folds. The graph across folds displays a consistent
¹⁴⁰⁷ upward trend for the training accuracy. However, there is an observable change in
¹⁴⁰⁸ the performance, particularly in Folds 1 and 2, where it shows a slight downward
¹⁴⁰⁹ trend in the validation accuracy.

4.2. DEEP LEARNING ANALYSIS

65

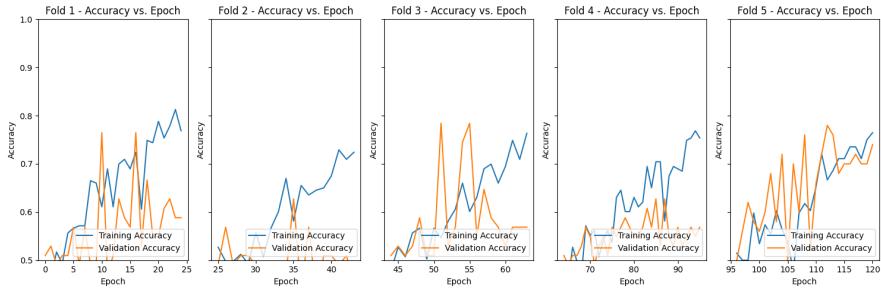


Figure 4.5: Training and validation accuracy per fold.

1410 Figure 4.6 shows the average performance of the model in terms of training and
 1411 validation accuracy across five folds. An upward trend is observable in the training
 1412 and validation accuracy, indicating that the model gradually improves over
 1413 the epochs. While fluctuations or dips can be seen in the validation accuracy,
 1414 the model recovers in later epochs. The training accuracy remains consistently
 1415 higher than the validation accuracy, which is expected behavior, as it learns from
 1416 the training data. Generally, the model demonstrates a gradual improvement in
 1417 learning, as reflected in the average upward trend aggregated across five folds.

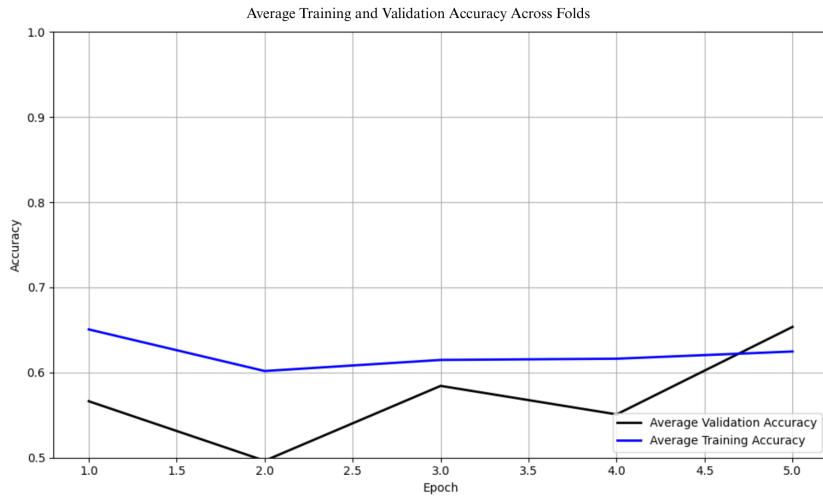


Figure 4.6: Average training and validation accuracy across folds.

1418 Figure 4.7 shows the performance of the model in the training and validation in

1419 terms of the training and validation loss across five folds. The graph across folds
 1420 displays a consistent downward trend for the training loss. On the other hand,
 1421 there is an observable change in the performance, especially in Folds 1,2,3, and 4,
 1422 where it shows an upward trend in the validation loss. This is an implication for
 1423 the learning performance of the model, as it may not be learning effectively.

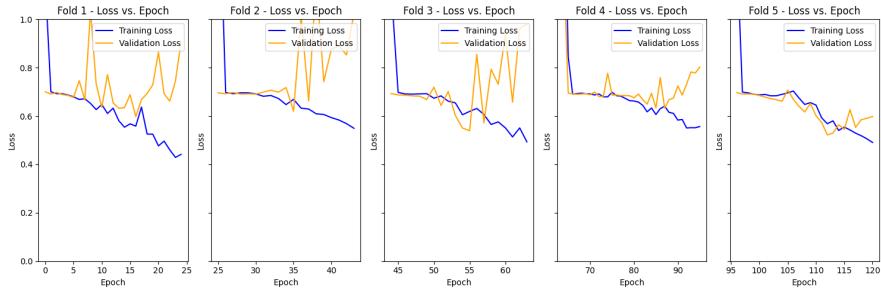


Figure 4.7: Training and validation loss per fold.

1424 Figure 4.8 shows the average performance of the model regarding training and
 1425 validation loss across five folds. A continuous downward trend is observed in
 1426 training and validation accuracy, indicating that the model's loss gradually de-
 1427 creases across epochs. This suggests that the model generalizes better following
 1428 the initial instability in the earlier epoch in the folds. Additionally, the training
 1429 loss consistently remains lower than the validation loss, since the model was di-
 1430 rectly optimized on the training set. Overall, the downward trend in training and
 1431 validation loss signifies that the model is learning and improving across the five
 1432 folds.

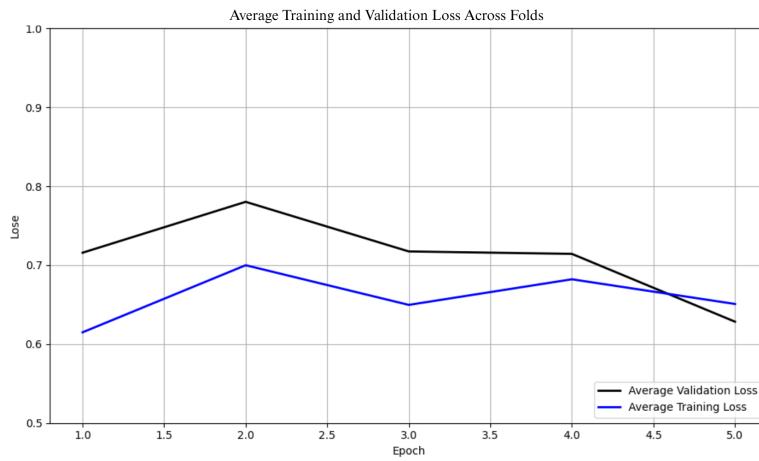


Figure 4.8: Average training and validation loss across folds.

1433 Figure 4.9 shows the confusion matrix for the true class label and predicted class
1434 label after the training and validation. The matrix shows the correctly predicted
1435 male and female samples and their corresponding percentages. Females have
1436 slightly higher true positives compared to males in the number and percentages,
1437 which are 94 and 88, corresponding to 74% and 69%, respectively. Additionally,
1438 the falsely classified samples were 33 for females and 39 for males, respectively,
1439 accounting for 26% and 31%.

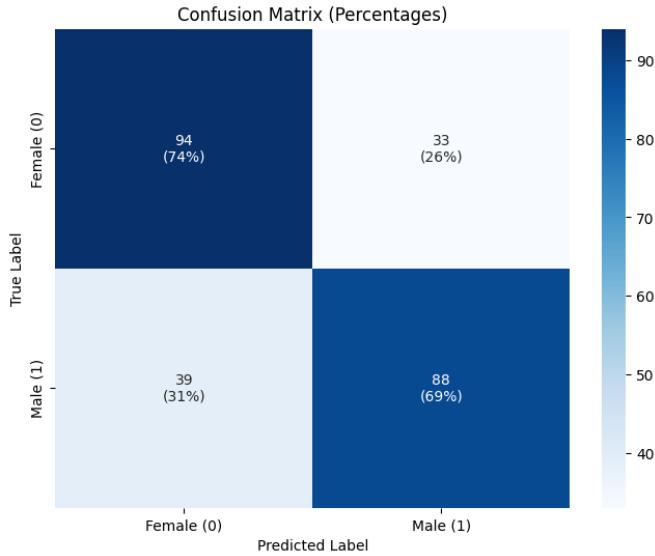


Figure 4.9: Confusion matrix for final model predictions.

1440 Figure 4.10 shows the average ROC curve, showing the proposed model's ability
 1441 to correctly identify the true positives, which can help determine the trade-off
 1442 between specificity and sensitivity. It will also determine the model's validity,
 1443 supporting that it is not being predicted based only on random chance. The
 1444 range of AUC ROC is between 0.5 and 1. The model achieved an average score
 1445 of 0.7734, which is better than random chance and a positive indication that the
 1446 model is performing reasonably.

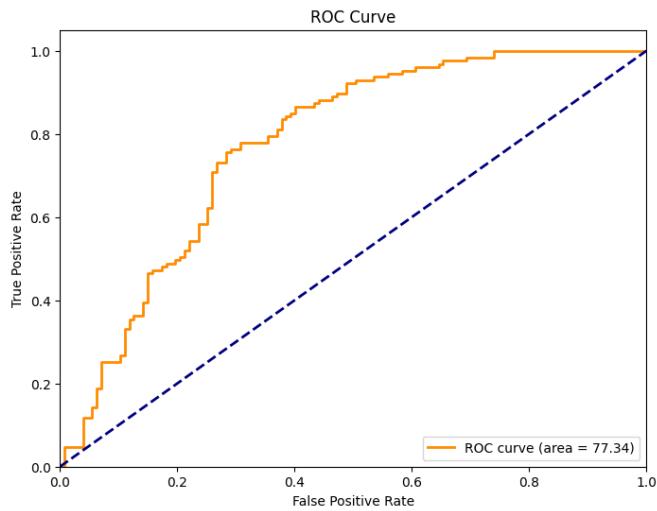


Figure 4.10: ROC curve with AUC score for the proposed model.

1447 4.3 Discussions

1448 This study aimed to develop a non-invasive method for identifying the sex of *T.*
 1449 *granosa* using machine learning and deep learning technologies. Specifically, it
 1450 explored the relevance of linear shell measurements and image data in building
 1451 accurate classification models that can support sustainable aquaculture practices.

1452 In the machine learning experiments, feature selection played a key role in enhanc-
 1453 ing model performance. A reduced set of five statistically significant features,
 1454 which were identified through Mann-Whitney U and Kruskal-Wallis tests, out-
 1455 performed models using all available features. The K-Nearest Neighbors (KNN)
 1456 classifier, trained on these five features, achieved an accuracy of 64.16%, precision
 1457 of 64.97%, recall of 64.16%, and an F1 score of 63.57%. The width-height ratio
 1458 observed from the left lateral view emerged as the most discriminative feature,
 1459 with a correlation score of $p = 0.18$.

1460 Deep learning experiments further revealed the impact of image angle and hyper-

¹⁴⁶¹ parameter tuning on classification performance. The left lateral view consistently
¹⁴⁶² yielded the highest metrics, with the best model reaching 71.68% accuracy, 72.52%
¹⁴⁶³ precision, 69.29% recall, 69.12% F1 score, and 77.34% AUC using a batch size of
¹⁴⁶⁴ 32 and 50 training epochs. Additionally, balanced dataset and activation function
¹⁴⁶⁵ contributed to improved model performance.

¹⁴⁶⁶ The improved accuracy from models using fewer, more relevant features supports
¹⁴⁶⁷ the idea that dimensionality reduction, when guided by statistical analysis, can
¹⁴⁶⁸ enhance classification. The prominence of the left lateral view in both machine
¹⁴⁶⁹ learning and deep learning results suggests that this angle reveals key morpho-
¹⁴⁷⁰ logical traits tied to sex differentiation. This aligns with the biological premise
¹⁴⁷¹ that some external characteristics may be more distinguishable when viewed from
¹⁴⁷² specific perspectives.

¹⁴⁷³ These findings demonstrates the feasibility of a non-invasive, accurate, and scal-
¹⁴⁷⁴ able approach to sex identification in *T. granosa*. This is especially important
¹⁴⁷⁵ in aquaculture, where traditional sex identification methods are often invasive or
¹⁴⁷⁶ require specialized knowledge. By reducing the need for physical intervention, this
¹⁴⁷⁷ approach promotes animal welfare and operational efficiency, potentially enabling
¹⁴⁷⁸ real-time sex identification in aquaculture settings.

¹⁴⁷⁹ When compared to related work, such as the gender classification of Chinese mit-
¹⁴⁸⁰ ten crabs using CNNs (Cui et al., 2020), this study reflects both shared method-
¹⁴⁸¹ ologies and important distinctions. While both utilized CNN architectures, differ-
¹⁴⁸² ences in image resolution, dataset characteristics, and species-specific morphology
¹⁴⁸³ may explain the performance gap which is 98.90% in the crab study compared to
¹⁴⁸⁴ 71.68% in this study. The lower accuracy here likely reflects subtler morphological

¹⁴⁸⁵ differences in *T. granosa* and limited dataset size.

¹⁴⁸⁶ Despite promising results, the study has several limitations. The dataset size (271
¹⁴⁸⁷ samples) was relatively small, which may have affected model generalizability.

¹⁴⁸⁸ Additionally, image data was constrained to six fixed angles, potentially missing
¹⁴⁸⁹ other informative views. These limitations may restrict the model's effectiveness
¹⁴⁹⁰ across diverse populations or environmental conditions.

¹⁴⁹¹ Chapter 5

¹⁴⁹² Conclusion and

¹⁴⁹³ Recommendations

¹⁴⁹⁴ 5.1 Conclusion

¹⁴⁹⁵ This study aimed to develop a non-invasive approach for sex identification of *Te-*
¹⁴⁹⁶ *gillarca granosa* using morphometric and morphological characteristics through
¹⁴⁹⁷ the integration of machine learning and deep learning technologies. In particu-
¹⁴⁹⁸ lar, it sought to determine whether measurable shell features and image-based
¹⁴⁹⁹ characteristics could reliably distinguish between male and female blood cockles.

¹⁵⁰⁰ The findings support the feasibility of this approach, with the proposed CNN
¹⁵⁰¹ model achieving a classification accuracy of 71.68%. This performance demon-
¹⁵⁰² strates that morphological and linear features, when processed through deep
¹⁵⁰³ learning, can serve as reliable indicators of sex in *T. granosa*. In comparison

1504 to traditional, more invasive methods such as dissection or spawning observation,
1505 this method presents a promising alternative that could be useful in aquaculture
1506 operations requiring rapid and non-destructive sex identification.

1507 The study also contributes a manually curated dataset of labeled images and shell
1508 measurements, which can serve as a foundation for further studies in this underex-
1509 plored domain. By emphasizing non-invasiveness, the research addresses a crucial
1510 need in sustainable aquaculture practices, particularly in improving broodstock
1511 selection without harming specimens.

1512 Although challenges such as limited sample size and computing resources were
1513 encountered, the overall results suggest that machine learning and deep learning
1514 techniques offer a scalable and practical solution for this biological classification
1515 task. As such, the study lays the groundwork for future research to expand the
1516 dataset, explore more advanced neural architectures, and develop real-time sex
1517 identification systems suitable for field or hatchery deployment.

1518 5.2 Recommendations

1519 This special problem aims to serve as a foundational study for future work involv-
1520 ing the application of machine learning and deep learning in aquaculture. Given
1521 the importance of accurate sex identification for breeding and stock management,
1522 several recommendations are proposed to enhance future studies.

1523 Future work should consider incorporating shape analysis and exploring more ad-
1524 vanced deep learning architectures, such as ResNet, SqueezeNet, and Inception-

1525 Net. The use of transfer learning may also enhance classification performance,
1526 especially when working with limited datasets. Real-time sex identification could
1527 be achieved by developing a system that captures rotational views of the shell
1528 from dorsal, lateral, and anterior angles.

1529 Due to time constraints, this study utilized a dataset of 1,626 images, with 271 im-
1530 ages per angle. Increasing the number and diversity of samples can help improve
1531 model generalization and robustness. Expanding the dataset to include differ-
1532 ent populations and environmental contexts would provide a more comprehensive
1533 understanding of morphological variation in *T. granosa*.

1534 Moreover, future researchers are encouraged to establish a controlled image ac-
1535 quisition environment, using a green or neutral background, consistent lighting,
1536 and fixed camera positioning. Image processing techniques, such as morphologi-
1537 cal transformations or background removal, can be employed to highlight relevant
1538 features and enhance model accuracy.

1539 The dataset produced in this study may serve as a valuable resource for future
1540 research in deep learning and marine biology. It can be further analyzed using ad-
1541 vanced techniques to uncover patterns of sexual dimorphism and develop scalable,
1542 real-time applications for aquaculture settings.

₁₅₄₃ **Chapter 6**

₁₅₄₄ **References**

- ₁₅₄₅ Adams, D. C., Rohlf, F. J., & Slice, D. E. (2004). Geometric morphometrics: ten years of progress following the ‘revolution’. *Italian Journal of Zoology*, *71*, 5–16. doi: 10.1080/11250000409356545
- ₁₅₄₆
- ₁₅₄₇
- ₁₅₄₈ Afifiati, N. (2007, 01). Gonad maturation of two intertidal blood clams *Anadara granosa* (l.) and *Anadara antiquata* (l.) (bivalvia: Arcidae) in central java. *10*.
- ₁₅₄₉
- ₁₅₅₀
- ₁₅₅₁ Aji, L. P. (2011). Review: Spawning induction in bivalve. *Jurnal Penelitian Sains*, *14*.
- ₁₅₅₂
- ₁₅₅₃ Arifin, W. A., Ariawan, I., Rosalia, A. A., Lukman, L., & Tufailah, N. (2021). Data scaling performance on various machine learning algorithms to identify abalone sex. *Jurnal Teknologi Dan Sistem Komputer*, *10*(1), 26–31. doi: 10.14710/jtsiskom.2021.14105
- ₁₅₅₄
- ₁₅₅₅
- ₁₅₅₆
- ₁₅₅₇ Arkhipkin, A. I. (2005). Statoliths as ‘black boxes’ (life recorders) in squid. *Marine and Freshwater Research*, *56*, 573–583. doi: 10.1071/mf04158
- ₁₅₅₈
- ₁₅₅₉ Awan, A. A. (2022, November). *A complete guide to data augmentation*.

- 1560 *tion.* Retrieved from [https://www.datacamp.com/tutorial/complete
1561 -guide-data-augmentation](https://www.datacamp.com/tutorial/complete-guide-data-augmentation)
- 1562 Aypa, S. M., & Baconguis, S. R. (2000). Philippines: Mangrove-friendly aquacul-
1563 ture. In J. H. Primavera, L. M. B. Garcia, M. T. Castaños, & M. B. Sur-
1564 tida (Eds.), *Mangrove-friendly aquaculture: Proceedings of the workshop on
1565 mangrove-friendly aquaculture organized by the SEAFDEC aquaculture de-
1566 partment, january 11-15, 1999, Iloilo City, Philippines* (pp. 41–56).
- 1567 BFAR. (2019). *Philippine fisheries profile 2018* (Tech. Rep.). PCA Compound,
1568 Elliptical Road, Quezon City, Philippines: Bureau of Fisheries and Aquatic
1569 Resources.
- 1570 Boey, P.-L., Maniam, G. P., Hamid, S. A., & Ali, D. M. H. (2011). Utilization
1571 of waste cockle shell (*Anadara granosa*) in biodiesel production from palm
1572 olein: Optimization using response surface methodology. *Fuel*, 90(7), 2353–
1573 2358. doi: 10.1016/j.fuel.2011.03.002
- 1574 Breton, S., Capt, C., Guerra, D., & Stewart, D. (2017, June). *Sex determining
1575 mechanisms in bivalves*. Preprints.org. doi: 10.20944/preprints201706.0127
1576 .v1
- 1577 Breton, S., Stewart, D. T., Shepardson, S., Trdan, R. J., Bogan, A. E., Chapman,
1578 E. G., ... Hoeh, W. R. (2010). Novel protein genes in animal mtDNA: A
1579 new sex determination system in freshwater mussels (bivalvia: Unionoida)?
1580 *Molecular Biology and Evolution*, 28(5), 1645–1659. doi: 10.1093/molbev/
1581 msq345
- 1582 Budd, A., Banh, Q., Domingos, J., & Jerry, D. (2015). Sex control in fish: Ap-
1583 proaches, challenges and opportunities for aquaculture. *Journal of Marine
1584 Science and Engineering*, 3(2), 329–355. doi: 10.3390/jmse3020329
- 1585 Burdon, D., Callaway, R., Elliott, M., Smith, T., & Wither, A. (2014, 04).

- 1586 Mass mortalities in bivalve populations: A review of the edible cockle
1587 *Cerastoderma edule* (l.). *Estuarine, Coastal and Shelf Science*, 150. doi:
1588 10.1016/j.ecss.2014.04.011
- 1589 Campos, A., Tedesco, S., Vasconcelos, V., & Cristobal, S. (2012). Proteomic
1590 research in bivalves: Towards the identification of molecular markers of
1591 aquatic pollution. *Proteomic Research in Bivalves*, 75(14), 4346–4359. doi:
1592 10.1016/j.jprot.2012.04.027
- 1593 Cerqueira, V., Santos, M., Baghoussi, Y., & Soares, C. (2024). On-the-fly data
1594 augmentation for forecasting with deep learning. *ArXiv (Cornell University)*. Retrieved from <https://doi.org/10.48550/arxiv.2404.16918>
- 1595 1596 Coe, W. R. (1943). Sexual differentiation in mollusks. i. pelecypods. *The Quarterly
Review of Biology*, 18, 154–164. doi: 10.1086/qrb.1943.18.issue-2
- 1597 1598 Collin, R. (2013). Phylogenetic patterns and phenotypic plasticity of molluscan
1599 sexual systems. *Integrative and Comparative Biology*, 53, 723–735. doi:
1600 10.1093/icb/ict076
- 1601 Concepcion, R., Guillermo, M., Tanner, S. E., Fonseca, V., & Duarte, B. (2023).
1602 Bivalvenet: A hybrid deep neural network for common cockle (*Cerastoderma
edule*) geographical traceability based on shell image analysis. *Ecological
Informatics*, 78, 102344. doi: 10.1016/j.ecoinf.2023.102344
- 1603 1604 Cui, Y., Pan, T., Chen, S., & Zou, X. (2020). A gender classification method
1605 for chinese mitten crab using deep convolutional neural network. *Multi-
1606 media Tools and Applications*, 79(11-12), 7669–7684. doi: [https://doi.org/
1607 1608 10.1007/s11042-019-08355-w](https://doi.org/10.1007/s11042-019-08355-w)
- 1609 Davenport, J., & Wong, T. (1986, September). Responses of the blood cockle
1610 *Anadara granosa* (l.) (bivalvia: Arcidae) to salinity, hypoxia and aerial ex-
posure. *Aquaculture*, 56(2), 151–162. Retrieved from <https://doi.org/>
- 1611

- 1612 10.1016/0044-8486(86)90024-4 doi: 10.1016/0044-8486(86)90024-4
- 1613 Doering, P., & Ludwig, J. (1990). Shape analysis of otoliths—a tool for indirect
1614 ageing of eel, *Anguilla anguilla* (L.)? *International Review of Hydrobiology*,
1615 75(6), 737–743. doi: 10.1002/iroh.19900750607
- 1616 Erica, D. (2018, April 4). *Clam dissection: A first step into dissection and*
1617 *anatomy for young learners*. Rosie Research. Retrieved from <https://rosieresearch.com/clam-dissection-anatomy/>
- 1619 Fao 2024 report: Sustainable aquatic food systems important for global food
1620 security – european fishmeal. (2024). <https://effop.org/news-events/fao-2024-report-sustainable-aquatic-food-systems-important-for-global-food-security/>.
- 1623 Ferguson, G. J., Ward, T. M., & Gillanders, B. M. (2011). Otolith shape and
1624 elemental composition: Complementary tools for stock discrimination of
1625 mulloway (*Argyrosomus japonicus*) in southern australia. *Fish Research*,
1626 110, 75–83. doi: 10.1016/j.fishres.2011.03.014
- 1627 GeeksforGeeks. (2020, August 6). *Stratified k fold cross validation*.
1628 Retrieved from <https://www.geeksforgeeks.org/stratified-k-fold-cross-validation/>
- 1630 GeeksforGeeks. (2024, May). *Cross-validation using k-fold with scikit-learn*. Re-
1631 tried from <https://www.geeksforgeeks.org/cross-validation-using-k-fold-with-scikit-learn/> (Accessed: 2025-04-23)
- 1633 Gosling, E. (2004). *Bivalve molluscs: biology, ecology and culture*. Oxford: Black-
1634 well Science.
- 1635 Heller, J. (1993). Hermaphroditism in molluscs. *Biological Journal of the Linnean*
1636 *Society*, 48, 19–42. doi: 10.1111/bij.1993.48.issue-1
- 1637 Hussain, S. S., & Zaidi, S. S. H. (2024). Adaboost ensemble approach with

- 1638 weak classifiers for gear fault diagnosis and prognosis in dc motors. *Applied
1639 Sciences*, 14(7), 3105. doi: 10.3390/app14073105
- 1640 Ishak, A. R., Mohamad, S., Soo, T. K., & Hamid, F. S. (2016). Leachate and
1641 surface water characterization and heavy metal health risk on cockles in
1642 kuala selangor. In *Procedia - social and behavioral sciences* (Vol. 222, pp.
1643 263–271). doi: 10.1016/j.sbspro.2016.05.156
- 1644 Jaiswal, S. (2024, January 4). *What is normalization in machine learning? a com-
1645 prehensive guide to data rescaling*. DataCamp. Retrieved from [https://www
1647 .datacamp.com/tutorial/normalization-in-machine-learning](https://www
1646 .datacamp.com/tutorial/normalization-in-machine-learning) (Accessed 2025-04-23)
- 1648 Karapunar, B., Werner, W., Fürsich, F. T., & Nützel, A. (2021). The ear-
1649 liest example of sexual dimorphism in bivalves—evidence from the astar-
1650 tid *Nicaniella* (lower jurassic, southern germany). *Journal of Paleontology*,
1651 95(6), 1216–1225. doi: 10.1017/jpa.2021.48
- 1652 Kerr, L. A., & Campana, S. E. (2014). Chemical composition of fish hard parts
1653 as a natural marker of fish stocks. In *Stock identification methods* (pp. 205–
1654 234). Elsevier. doi: 10.1016/b978-0-12-397003-9.00011-4
- 1655 Kim, E., Yang, S.-M., Cha, J.-E., Jung, D.-H., & Kim, H.-Y. (2024). Deep
1656 learning-based phenotype classification of three ark shells: *Anadara kagoshi-*
1657 *mensis*, *Tegillarca granosa*, and *Anadara Broughtonii*. *Frontiers in Marine
1658 Science*, 11. doi: 10.3389/fmars.2024.1356356
- 1659 Lee, J. H. (1997). Studies on the gonadal development and gametogenesis of the
1660 granulated ark, *Tegillarca granosa* (linne). *The Korean Journal of Malacol-
1661 ogy*, 13, 55–64.
- 1662 Leguá, J., Plaza, G., Pérez, D., & Arkhipkin, A. (2013). Otolith shape analysis as
1663 a tool for stock identification of the southern blue whiting, *Micromesistius*

- 1664 *australis*. *Latin American Journal of Aquatic Research*, 41, 479–489.
- 1665 Mahé, K., Oudard, C., Mille, T., Keating, J., Gonçalves, P., Clausen, L. W., &
1666 et al. (2016). Identifying blue whiting (*Micromesistius poutassou*) stock
1667 structure in the northeast atlantic by otolith shape analysis. *Canadian*
1668 *Journal of Fisheries and Aquatic Sciences*, 73, 1363–1371. doi: 10.1139/
1669 cjfas-2015-0332
- 1670 May, K., Maung, C., Phy, E., & Tun, N. (2021). Spawning period of blood cockle
1671 *Tegillarca granosa* (linnaeus, 1758) in myeik coastal areas. *J. Myanmar*
1672 *Acad. Arts Sci*, 4.
- 1673 Miranda, D. V., & Ferriols, V. M. E. N. (2023). Initial attempts on spawning
1674 and larval rearing of the blood cockle, *Tegillarca granosa* (linnaeus, 1758),
1675 in the philippines. *Asian Fisheries Science*, 36(2). doi: 10.33997/j.afs.2023
1676 .36.2.001
- 1677 Mérigot, B., Letourneur, Y., & Lecomte-Finiger, R. (2007). Characterization of
1678 local populations of the common sole *solea solea* (pisces, soleidae) in the nw
1679 mediterranean through otolith morphometrics and shape analysis. *Marine*
1680 *Biology*, 151(3), 997–1008. doi: 10.1007/s00227-006-0549-0
- 1681 Nahm, F. S. (2022). Receiver operating characteristic curve: Overview and
1682 practical use for clinicians. *Korean Journal of Anesthesiology*, 75(1), 25–
1683 36. Retrieved from <https://doi.org/10.4097/kja.21209> doi: 10.4097/
1684 kja.21209
- 1685 Narasimham, K. A. (1988). Taxonomy of the blood clams anadara (*Tegillarca*
1686 *granosa* (linnaeus, 1758) and a. (t.) *rhombea* (born, 1780).
- 1687 Naylor, R. L., Goldburg, R. J., Primavera, J. H., Kautsky, N., Beveridge,
1688 M. C. M., Clay, J., ... Troell, M. (2000). Effect of aquaculture on world
1689 fish supplies. *Nature*, 405(6790), 1017–1024. doi: 10.1038/35016500

- 1690 Ponton, D. (2006). Is geometric morphometrics efficient for comparing otolith
1691 shape of different fish species? *Journal of Morphology*, 267(7), 750–757.
1692 doi: 10.1002/jmor.10439
- 1693 Quenu, M., Trewick, S. A., Brescia, F., & Morgan-Richards, M. (2020). Geometric
1694 morphometrics and machine learning challenge currently accepted species
1695 limits of the land snail *placostylus* (pulmonata: Bothriembryontidae) on the
1696 isle of pines, new caledonia. *Journal of Molluscan Studies*, 86(1), 35–41.
1697 doi: 10.1093/mollus/eyz031
- 1698 Ribeiro, D. (2024, Aug). *Diogo ribeiro. data science. kruskal-wallis test.*
1699 Retrieved from https://diogoribeiro7.github.io/statistics/data%20analysis/kruskal_wallis/ (Accessed: 2025-04-23)
- 1700 1701 Sany, S. B. T., Hashim, R., Rezayi, M., Salleh, A., Rahman, M. A., Safari, O.,
1702 & Sasekumar, A. (2014). Human health risk of polycyclic aromatic hydro-
1703 carbons from consumption of blood cockle and exposure to contaminated
1704 sediments and water along the klang strait, malaysia. *Marine Pollution
1705 Bulletin*, 84(1-2), 268–279. doi: 10.1016/j.marpolbul.2014.05.004
- 1706 1707 Srisunont, C., Nobpakhun, Y., Yamalee, C., & Srisunont, T. (2020). Influence
1708 of seasonal variation and anthropogenic stress on blood cockle (*Tegillarca
1709 Granosa*) production potential. *Influence of Seasonal Variation and Anthro-
1710 pogenic Stress on Blood Cockle (*Tegillarca Granosa*) Production Potential,*
1711 44(2), 62–82.
- 1712 Tarca, A. L., Carey, V. J., Chen, X.-w., Romero, R., & Drăghici, S. (2007). Ma-
1713 chine learning and its applications to biology. *PLoS Computational Biology*,
1714 3(6), e116. doi: 10.1371/journal.pcbi.0030116
- 1715 Team, K. (n.d.). *Keras documentation: Reducelronplateau.* Retrieved from
https://keras.io/api/callbacks/reduce_lr_on_plateau/

- 1716 Thompson, R. J., Newell, R. I. E., Kennedy, V. S., & Mann, R. (1996). Repro-
1717 ductive process and early development. In V. S. Kennedy, R. I. E. Newell,
1718 & A. F. Eble (Eds.), *The eastern oyster crassostrea virginica* (pp. 335–370).
1719 College Park, MD: Maryland Sea Grant.
- 1720 Tsutsumi, M., Saito, N., Koyabu, D., & Furusawa, C. (2023). A deep learning
1721 approach for morphological feature extraction based on variational auto-
1722 encoder: An application to mandible shape. *Npj Systems Biology and Ap-*
1723 *plications*, 9(1), 1–12. doi: 10.1038/s41540-023-00293-6
- 1724 Wong, T. M., & Lim, T. G. (2018). *Cockle (Anadara granosa) seed produced in*
1725 *the laboratory, malaysia*. (Handle.net) doi: 10.3366/in_3366.pdf
- 1726 Zahn, C. T., & Roskies, R. Z. (1972). Fourier descriptors for plane closed curves.
1727 *IEEE Transactions on Computers*, C-21, 269–281. doi: 10.1109/tc.1972
1728 .5008949
- 1729 Zelditch, M., Swiderski, D. L., & Sheets, H. D. (2004). *Geometric morphometrics*
1730 *for biologists: A primer* (2nd ed.). Waltham: Elsevier Academic Press.
- 1731 Zha, S., Tang, Y., Shi, W., Liu, H., Sun, C., Bao, Y., & Liu, G. (2022). Im-
1732 pacts of four commonly used nanoparticles on the metabolism of a ma-
1733 rine bivalve species, *Tegillarca granosa*. *Chemosphere*, 296, 134079. doi:
1734 10.1016/j.chemosphere.2022.134079
- 1735 Zhan, P. L., Zha, & Bao, Y. (2022). Hypoxia-mediated immunotoxicity in the
1736 blood clam *Tegillarca granosa*. *Marine Environmental Research*, 177. Re-
1737 trieved from <https://doi.org/10.1016/j.marenvres.2022.105632>

¹⁷³⁸ **Appendix A**

¹⁷³⁹ **Code Snippets**

¹⁷⁴⁰ **i. Machine Learning**

¹⁷⁴¹ This section displays the key steps in the machine learning analysis by performing
¹⁷⁴² feature engineering to create and transform a new dataset, identifying the most
¹⁷⁴³ significant features through the Kruskal-Wallis Test, applying random undersam-
¹⁷⁴⁴ pling to address the minimal imbalance in the dataset, and conducting five-fold
¹⁷⁴⁵ cross-validation to evaluate the model's performance.

```
female_litob['LW_ratio']= female_litob['Length'] / female_litob['Width']
female_litob['LH_ratio'] = female_litob['Length'] / female_litob['Height']
female_litob['WH_ratio'] = female_litob['Width'] / female_litob['Height']
# female_litob['DU_ratio'] = female_litob['Distance Umbos'] / female_litob['Height']
female_litob['UL_ratio'] = female_litob['Distance Umbos'] / female_litob['Length']
female_litob['HL_ratio'] = female_litob['Length (Hinge Line)'] / female_litob['Length']
female_litob['UH_ratio'] = female_litob['Distance Umbos'] / female_litob['Height']
female_litob['Rib Density'] = female_litob['Rib count'] / female_litob['Length']
```

Figure A.1: Feature engineering used to create and transform the dataset for machine learning analysis.

```

sorted_features = feature_importance_scores.sort_values(ascending=False)
colors = sns.color_palette("Blues_d", len(sorted_features))
colors = colors[::-1]
plt.figure(figsize=(10, 6))

# Map codenames to sorted_features.index
sorted_features.index = sorted_features.index.map(codenames)

sns.barplot(x=sorted_features.values, y=sorted_features.index, hue = sorted_features.index, palette=colors,dodge=False,leg
plt.xlabel("Kruskal-Wallis Statistic")
plt.ylabel("Features")
plt.title("Feature Importance based on Kruskal-Wallis Test")
plt.show()

```

Figure A.2: Feature importance scores derived from the Kruskal-Wallis test to identify the most significant variables.

```

from imblearn.under_sampling import RandomUnderSampler

rus = RandomUnderSampler(sampling_strategy=1) # Numerical value
# rus = RandomUnderSampler(sampling_strategy="not minority") # String
X_res, y_res = rus.fit_resample(X, y)

ax = y_res.value_counts().plot.pie(autopct='%.2f')
_ = ax.set_title("Under-sampling")

```

Figure A.3: Random undersampling applied in machine learning to address class imbalance.

1746 ii. Image Processing

1747 This section of the paper displays the key steps in the image processing by resizing
 1748 the images to have similar dimensions of 256x256, and the shadows were removed
 1749 to improve the image quality, and remove noise before proceeding to the deep
 1750 learning operations.

```

for model_name, (model, param_grid) in models_and_grids.items():
    print(f"Training {model_name}...")
    grid_search = GridSearchCV(estimator=model, param_grid=param_grid, cv=5, scoring='accuracy', return_train_score=False,
                               grid_search.fit(X, y)

    # Get the best estimator's CV results
    best_index = grid_search.best_index_

    # Get fold scores for the best parameters only
    fold_scores = [
        grid_search.cv_results_['split0_test_score'][best_index],
        grid_search.cv_results_['split1_test_score'][best_index],
        grid_search.cv_results_['split2_test_score'][best_index],
        grid_search.cv_results_['split3_test_score'][best_index],
        grid_search.cv_results_['split4_test_score'][best_index]
    ]

    # Calculate the average score across folds
    avg_score = sum(fold_scores) / len(fold_scores)

    # Convert to percentage by multiplying by 100 for display
    fold_scores_percentage = [score * 100 for score in fold_scores]
    avg_score_percentage = avg_score * 100

    # Round each fold score individually
    fold_scores_rounded = [round(score, 2) for score in fold_scores_percentage]

    # Round the average score
    avg_score_rounded = round(avg_score_percentage, 2)

    # Append the fold scores and average score to the list
    model_scores.append({
        'Model': model_name,
        'Fold 1': fold_scores_rounded[0],
        'Fold 2': fold_scores_rounded[1],
        'Fold 3': fold_scores_rounded[2],
        'Fold 4': fold_scores_rounded[3],
        'Fold 5': fold_scores_rounded[4],
        'Average CV Score': avg_score_rounded,
        'Best Parameters': grid_search.best_params_
    })
)

```

Figure A.4: Five-fold cross-validation used to evaluate and tune machine learning model performance.

```

# Process each image
for img_name in image_files:
    img_path = os.path.join(input_male, img_name)
    output_path = os.path.join(output_male, img_name)

    # Read the image
    image = cv2.imread(img_path)
    if image is None:
        print(f"Skipping invalid image: {img_path}")
        continue

    # Resize and pad the image to 256x256
    resized_image = resize_and_pad(image, 256)

```

Figure A.5: Resizing images to 256x256 pixels for consistent input dimensions.

```

# Convert to HSV and apply threshold
frame_HSV = cv.cvtColor(frame, cv.COLOR_BGR2HSV)
frame_threshold = cv.inRange(frame_HSV, (low_H, low_S, low_V), (high_H, high_S, high_V))

# Filling holes
im_floodfill = frame_threshold.copy()
h, w = frame_threshold.shape[:2]
mask = np.zeros((h+2, w+2), np.uint8)
cv.floodFill(im_floodfill, mask, (0, 0), 255)
im_floodfill_inv = cv.bitwise_not(im_floodfill)
mask = frame_threshold | im_floodfill_inv

# Apply morphological operations
kernel = np.ones((3, 3), np.uint8)
mask = cv.morphologyEx(mask, cv.MORPH_OPEN, kernel, iterations=2)
mask = cv.morphologyEx(mask, cv.MORPH_CLOSE, kernel, iterations=4)

# Find contours
contours, _ = cv.findContours(mask, cv.RETR_EXTERNAL, cv.CHAIN_APPROX_SIMPLE)

if contours:
    # Merge contours using convex hull
    hull = cv.convexHull(np.vstack(contours))

    # Create a mask for the shell
    shell_mask = np.zeros_like(frame)
    cv.drawContours(shell_mask, [hull], -1, (255, 255, 255), -1)

    # Create a white background
    white_background = np.ones_like(frame) * 255

    # Combine the shell with the white background
    result = np.where(shell_mask == 255, frame, white_background)
else:
    result = frame # If no contour is found, return the original image

# Save the processed image
cv.imwrite(output_path, result)

```

Figure A.6: Processing the images to remove the shadows.

1751 **iii. Deep Learning**

1752 This section of the paper displays the key steps in deep learning by implementing
 1753 random undersampling in addressing class imbalance, data augmentation to create
 1754 variability in the dataset, and the CNN layers comprised of three convolution
 1755 layers, a flatten, and 2 dense layers.

```
# Get male and female filenames
male_samples = sorted(os.listdir(male_folder))
female_samples = sorted(os.listdir(female_folder))

# Randomly sample 127 male samples to match female sample size
male_samples_to_keep = random.sample(male_samples, undersample_size)

# Copy the selected male samples to the balanced male directory
for file in male_samples_to_keep:
    shutil.copy(os.path.join(male_folder, file), os.path.join(balanced_male_dir, file))

# Copy all female samples to the balanced female directory (since it's already balanced)
for file in female_samples:
    shutil.copy(os.path.join(female_folder, file), os.path.join(balanced_female_dir, file))
```

Figure A.7: Random undersampling applied in deep learning to balance class distribution in the datasets.

```
def create_data_augmentation():
    return tf.keras.Sequential([
        layers.RandomFlip("horizontal"),
        layers.RandomRotation(0.05),
        layers.RandomZoom(0.05),
    ])
```

Figure A.8: On-the-fly data augmentation used to create variety of random transformation to increase training images.

```
def create_cnn_model(img_width=256, img_height=256):
    model = Sequential([
        layers.Input(shape=(img_width, img_height, 3)),
        layers.Rescaling(1./255),
        layers.Conv2D(16, (3,3), activation='relu'),
        layers.MaxPooling2D(2,2),
        layers.Conv2D(32, (3,3), activation='relu'),
        layers.MaxPooling2D(2,2),
        layers.Conv2D(64, (3,3), activation='relu'),
        layers.MaxPooling2D(2,2),
        layers.Flatten(),
        layers.Dense(128, activation='relu'),
        layers.Dropout(0.5),
        layers.Dense(1, activation='sigmoid')
    ])
    return model
```

Figure A.9: CNN architecture used for training the image dataset.

¹⁷⁵⁶ **Appendix B**

¹⁷⁵⁷ **Resource Persons**

¹⁷⁵⁸ This section of the paper presents information about the resource persons who
¹⁷⁵⁹ contributed to and assisted the researchers during the data gathering process.

¹⁷⁶⁰ **Dr. Victor Marco Emmanuel N. Ferriols**

¹⁷⁶¹ Provided blood cockles samples used in this study

¹⁷⁶² Director, University of the Philippines Institute of Aquaculture

¹⁷⁶³ vnferriols@up.edu.ph

¹⁷⁶⁴

¹⁷⁶⁵ **Ms. Allena Esther D. Arteta**

¹⁷⁶⁶ Performed spawning of blood cockles samples, assisted the researchers with dis-
¹⁷⁶⁷ section and sex identification.

¹⁷⁶⁸ Research Associate, Institute of Aquaculture

¹⁷⁶⁹ adarteta@up.edu.ph

¹⁷⁷⁰

1771 Ms. LC May C. Gasit

- 1772 Performed spawning of blood cockles samples, assisted the researchers with the
1773 dissection and sex identification
1774 Research Associate, Institute of Aquaculture
1775 lcgasit@up.edu.ph

1776

1777 Sheila G. Untalan

- 1778 Performed spawning of blood cockles samples, assisted the researchers with the
1779 dissection and sex identification
1780 Research Associate, Institute of Aquaculture

1781

1782 Joel M. Fabrigas

- 1783 Assisted the researchers with the dissection and sex identification
1784 Hatchery Staff, Institute of Aquaculture

1785

1786 Paul Andre M. Lopez

- 1787 Assisted the researchers with the dissection and sex identification
1788 Hatchery Staff, Institute of Aquaculture

1789

¹⁷⁹⁰ **Appendix C**

¹⁷⁹¹ **Data Gathering Documentation**

¹⁷⁹² This section of the paper presents the data gathering process, including spawning,
¹⁷⁹³ dissection, sex identification, collection of linear measurements, and image capture
¹⁷⁹⁴ from six different camera angles.



Figure C.1: Sex identification through spawning of *T. granosa*.



Figure C.2: Sex-based separation of *T. granosa* samples post-spawning.



Figure C.3: Sex identified female through dissection of *T. granosa*.



Figure C.4: Sex identified male through dissection of *T. granosa*.

Litob_Id	Length	Width	Height	Rib count	Length (Hinge Line)	Distance Umbos
10001	48.05	37.6	32.15	20	33.55	4.1
20001	48.05	37.6	32.15	20	33.55	4.1
30001	48.05	37.6	32.15	20	33.55	4.1
40001	48.05	37.6	32.15	20	33.55	4.1
50001	48.05	37.6	32.15	20	33.55	4.1
60001	48.05	37.6	32.15	20	33.55	4.1
10002	47.4	32.5	32.25	20	33.1	3.05
20002	47.4	32.5	32.25	20	33.1	3.05
30002	47.4	32.5	32.25	20	33.1	3.05
40002	47.4	32.5	32.25	20	33.1	3.05
50002	47.4	32.5	32.25	20	33.1	3.05
60002	47.4	32.5	32.25	20	33.1	3.05
10003	43.3	34.1	31.25	21	32.05	4.5
20003	43.3	34.1	31.25	21	32.05	4.5
30003	43.3	34.1	31.25	21	32.05	4.5
40003	43.3	34.1	31.25	21	32.05	4.5
50003	43.3	34.1	31.25	21	32.05	4.5
60003	43.3	34.1	31.25	21	32.05	4.5
10075	50.05	35.05	32.05	21	30.05	4.1
20075	50.05	35.05	32.05	21	30.05	4.1

Figure C.5: Linear measurements of female *T. granosa*.

Litob_Id	Length	Width	Height	Rib count	Length (Hinge Line)	Distance Umbos
110004	43.1	33.05	28.15	21	28.5	3.05
120004	43.1	33.05	28.15	21	28.5	3.05
130004	43.1	33.05	28.15	21	28.5	3.05
140004	43.1	33.05	28.15	21	28.5	3.05
150004	43.1	33.05	28.15	21	28.5	3.05
160004	43.1	33.05	28.15	21	28.5	3.05
110005	41.1	31.05	27.6	20	23.05	3.35
120005	41.1	31.05	27.6	20	23.05	3.35
130005	41.1	31.05	27.6	20	23.05	3.35
140005	41.1	31.05	27.6	20	23.05	3.35
150005	41.1	31.05	27.6	20	23.05	3.35
160005	41.1	31.05	27.6	20	23.05	3.35
110006	43.2	33.45	29.35	20	29.35	3.3
120006	43.2	33.45	29.35	20	29.35	3.3
130006	43.2	33.45	29.35	20	29.35	3.3
140006	43.2	33.45	29.35	20	29.35	3.3
150006	43.2	33.45	29.35	20	29.35	3.3
160006	43.2	33.45	29.35	20	29.35	3.3
110007	41.5	32.55	27.7	20	24.1	3.7
120007	41.5	32.55	27.7	20	24.1	3.7

Figure C.6: Linear measurements of male *T. granosa*.

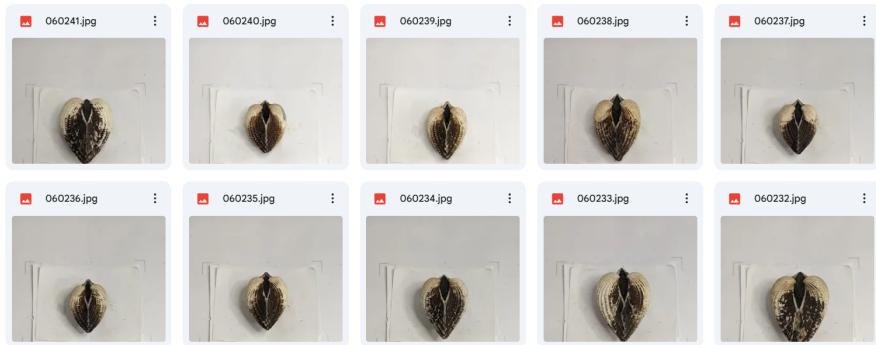


Figure C.7: Captured images of female *T. granosa*.

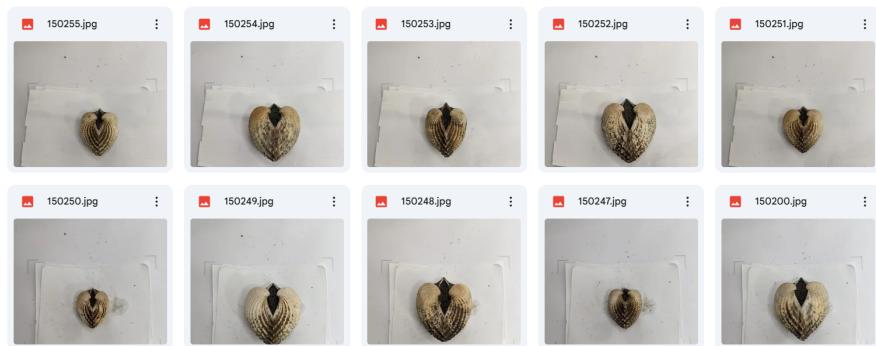


Figure C.8: Captured images of male *T. granosa*.