Maynooth University
National University of Ireland Maynooth

Statistical modelling and machine vision applied to automating animal monitoring systems

A dissertation submitted for the degree of
Doctor of Philosophy

By:

Gabriel Rodrigues Palma

Under the supervision of:

Prof. Rafael A. Moral

Dr. Charles Markham

Hamilton Institute
National University of Ireland Maynooth
Ollscoil na hÉireann, Má Nuad

July 2025

To my family, and friends that supported me with unconditional love.

Declaration

I hereby declare that I have produced this manuscript without the prohibited assistance of any third parties and without making use of aids other than those specified.

The thesis work was conducted from September 2021 to July 2025 under the supervision of Dr. Rafael A. Moral and Dr. Charles Markham in the Hamilton Institute, National University of Ireland Maynooth.

Gabriel Rodrigues Palma.

Maynooth, Ireland,

July 2025.

Sponsor

This publication has emanated from research conducted with the financial support of Taighde Éireann – Research Ireland under Grant number 18/CRT/6049.

SFI Centre for Research Training
in Foundations of Data Science

Collaborations

Rafael A. Moral: As my supervisor, Prof. Moral (Maynooth University) supervised and collaborated on the work of all chapters.

Charles Markham: As my joint-supervisor, Dr. Markham (Maynooth University) supervised and collaborated on the work of all chapters.

Rocío Alaiz: Dr. Rocío Alaiz (University of Leon) collaborated on the third section of Chapter 2 by providing insights to the proposal of methods to deal with scarce and class imbalance and reviewing the manuscript submitted to the Journal of Artificial Intelligence in Agriculture.

Rodrigo F. Mello: Dr. Rodrigo F. Mello (Mercado Livre) collaborated on the second section of Chapter 3 by providing insights on the development of the proposed approach and reviewing the paper published at Ecological Informatics.

Wesley A.C. Godoy: Dr. Wesley A.C. Godoy (University of São Paulo) collaborated on both sections of Chapter 3 by providing insights to the proposal of methods to predict insect outbreak and reviewing the papers published at Ecological Informatics.

Luciano M. Verdade: Dr. Luciano M. Verdade (University of São Paulo) collaborated on both the first section of Chapter 2 by collecting the data used in the case studies and reviewing the paper published at IMVIP.

Oliver Mason: Prof. Oliver Mason (Maynooth University) collaborated on the first section of Chapter 3 by brining insights towards the implementation of design of the Pattern-based prediction method and reviewing the paper published at Ecological Informatics.

Eduardo Engel: Eduardo Engel (University of São Paulo) collaborated on both sections of Chapter 3 by collecting the data used in the case studies and reviewing the paper published at Ecological Informatics.

Douglas Lau: Dr. Douglas Lau (Brazilian Agricultural Research Corporation, Embrapa Trigo) collaborated on both sections of Chapter 3 by collecting the data used in the case studies and reviewing the paper published at Ecological Informatics.

Ana Carla Aquino: Dr. Ana Carla Aquino (University of São Paulo) collaborated on both the first section of Chapter 2 by collecting the data used in the case studies and reviewing the paper published at IMVIP.

Patrícia F Monticelli: Dr. Patrícia F Monticelli (University of São Paulo) collaborated on both the first section of Chapter 2 by collecting the data used in the case studies and reviewing the paper published at IMVIP.

Alexandre S. Araújo: Alexandre S. Araújo (University of São Paulo) collaborated on the third section of Chapter 2 by collecting the data used in the case studies and reviewing the manuscript submitted to the Journal of Artificial Intelligence in Agriculture.

Marcoandre Savaris: Dr. Marcoandre Savaris (University of São Paulo) collaborated on the third section of Chapter 2 by by collecting the data used in the case studies and reviewing the manuscript submitted to the Journal of Artificial Intelligence in Agriculture.

Roberto A. Zucchi: Prof. Roberto A. Zucchi (University of São Paulo) collaborated on the third section of Chapter 2 by by collecting the data used in the case studies and reviewing the manuscript submitted to the Journal of Artificial Intelligence in Agriculture.

Conor Hackett: Conor Hackett (Maynooth University) collaborated on the second section of Chapter 2 by developing software and developing the experiments on Platforms of Computing.

Edgar Galvan: Dr. Edgar Galvan (Maynooth University) collaborated on the first section of Chapter 3 by bringing insights towards the implementation of the Differential evolution algorithm, the optimisation process of the Pattern-based prediction method and reviewing the paper published at Ecological Informatics.

Publications

The chapters in this thesis have been either published or submitted to peer-reviewed journals, books or conferences. The first section of Chapter 2 has been published in the proceedings of the Irish Machine Vision and Image Processing Conference, *IMVIP*, the second section of the same Chapter has been published by Springer as a chapter of the book *Modelling Insect Populations in Agricultural Landscapes*. The final section of Chapter 2 has been submitted to the Artificial Intelligence in Agriculture journal. Finally, both sections of Chapter 3 have been published in the journal *Ecological Informatics*.

Peer-reviewed journal articles:

- **Palma, G. R.**, Mello, R. F., Godoy, W. A., Engel, E., Lau, D., Markham, C., Moral, R. A. (2024) Forecasting insect abundance using time series embedding and machine learning. *Ecological Informatics*, Volume 85, 102934. https://doi.org/10.1016/j.ecoinf.2024.102934

- **Palma, G.R.**, Godoy, W. A., Engel, E., Lau, D., Galvan, E., Mason, O., Markham, C., Moral, R. A. (2023) Pattern-based prediction of population outbreaks. *Ecological Informatics*, vol 77, 102220. https://doi.org/10.1016/j.ecoinf.2023.102220

Peer-reviewed book chapter:

- **Palma, G.R.**, Hackett, C.P., Markham, C. (2023). Machine Vision Applied to Entomology. In: A. Moral, R., Godoy, W.A. (eds) Modelling Insect Populations in Agricultural Landscapes. Entomology in Focus, vol 8. Springer, Cham. https://doi.org/10.1007/978-3-031-43098-5_9

Peer-reviewed conference article:

- **Palma, G.R.**, Aquino, A.C.M.M., Monticelli, P.F., Verdade, L.M., Markham, C., and Moral, R.A. Moral. A machine vision system for avian song classification with CNN's. In Proceedings of the Irish Machine Vision and Image Processing Conference (2022), pp. 64–71. https://iprcs.github.io/pdf/IMVIP2022_Proceedings.pdf

- **Palma, G.R.**, Mello, F., Godoy, W. A., Engel, E., Lau, D., Moral, R.A. (2023) Forecasting the abundance of agricultural pests: a new machine learning framework. 9th Channel Network Conference, Wageningen, Netherlands.

- **Palma, G.R.**, Mello, F., Godoy, W. A., Engel, E., Lau, D., Moral, R.A. (2023) Forecasting insect abundance using time series embedding and environmental covariates. 37th International Workshop on Statistical Modelling, Dortmund, Germany.

Submitted articles (under review):

- **Palma, G.R.**, Alaiz, R., Araújo, A. S., Savaris, M., Zucchi, R. A., Markham, C., Moral, R. A. Towards species' classification of the *Anastrepha pseudoparallela* group. Submitted to *Methods in Artificial Intelligence in Agriculture* (2025).

Contents

Abstract

Conservation biology has guided multiple applications that relate to how humans interact with wildlife. The work of Caughley [1994] describes the core directions for conservation biology by introducing the main options for human intervention in nature. Caughley [1994] proposed the following management actions: (i) increase depleted populations; (ii) decrease excessive populations; (iii) establish maximum sustainable yields; (iv) carry out monitoring programs without additional actions over stable populations. In this thesis, we proposed new approaches to improve the state-of-the-art quantitative methods applied to challenges on two of Caughley's management actions: 1) carry out monitoring programs without additional actions over stable populations and 2) decrease excessive populations.

For the first management action, we explored the challenge of identifying avian species in audio-based monitoring systems using transfer learning and Convolution Neural Networks. We also explored the challenge of accurate insect classification and localisation in image-based monitoring systems focusing on species with agronomical and forensic importance. We proposed a framework for combining computer vision and machine learning to identify these species and explored solutions for this action in small and imbalanced datasets. Finally, for the second management action, we introduced two new statistical machine learning approaches to predict insect outbreaks and abundance to aid decision-making applied to Integrated Pest Management. We proposed a new statistical machine learning method, Pattern-Based Prediction (PBP), to predict insect outbreaks and a new approach combining statistical machine learning, causal analysis and time series embedding to guide the selection of climate time series and their lags to build statistical machine learning forecasting methods.

Acknowledgements

I would like to thank my Mother, Father, and fiancé for all the love and emotional support during my PhD experience.

I am incredibly grateful for the opportunity to be part of the Center For Research Training in Foundations of Data Science, and I would like to thank all the staff involved in the program. I extend my acknowledgements and express my gratitude to Taighde Éireann – Research Ireland, Skillnet Ireland, the Enterprise Alliance Partners and the Host institutions, Maynooth University, University of Limerick, and University College Dublin. My career has transformed, thanks to every training opportunity, industry, and international placement I have participated in provided by the CRT.

I would like to thank my PhD supervisor, Prof. Rafael Moral, for the numerous opportunities he provided me to develop my skills as an independent researcher and lecturer in statistics. The scientific challenges you presented to me were pivotal to my development as a researcher, and I am grateful for your guidance in preparing me for the following stages of my career. Moreover, thank you for the time you invested in me. We have been working together since 2018, and every interaction was a learning experience for me. I am honoured to have the opportunity to work with you and continue collaborating on future research projects.

I would like to thank my joint-supervisor, Dr. Charles Markham, for helping me develop my skills in machine vision and computer science and for providing me with opportunities to grow my independence in my research career continually. Every interaction I had with you was a valuable learning opportunity, and I appreciate your patience and ability to break down complex topics into straightforward and

programmable concepts. I learned a lot from you, and I am grateful that you accepted to work with me in 2020, as my passion for computer science has grown significantly since then, thanks to you. I am also honoured to have the opportunity to work with you and look forward to continuing our collaboration on other projects.

Finally, I express my gratitude to all the collaborators involved in this thesis. My experience with every collaboration was fantastic, and I thank all of you for the opportunity to learn from you through our pleasant collaborations. I want to thank Dr. Rocío Alaiz for the chance to work with you at the University of Leon. It was a fantastic experience, and I learned a lot working with you. So, I would like to express my gratitude for your hospitality and teachings during my visit to Spain.

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Motivation

Conservation biology has informed numerous applications related to human-wildlife interactions. The work of Caughley [1994] outlines the core directions for conservation biology by presenting the primary options for human intervention in nature. Caughley's proposed management actions involve (i) increasing depleted populations, (ii) decreasing excessive populations, (iii) establishing maximum sustainable yields, and (iv) carrying out monitoring programs without additional actions over stable populations [Caughley, 1994, Fryxell et al., 2014, Verdade et al., 2014, Krausman and Cain, 2022]. The conservation biologist community has accepted these actions as the core for developing wildlife management protocols [Krausman and Cain, 2022, Hone and Krebs, 2023, Martínez-Jauregui et al., 2023].

All of Caughley's management actions involve animal monitoring [Caughley, 1994]. Monitoring animal abundance provides insights that can be further utilised for data-driven decisions about the studied population [Wägele et al., 2022, Besson et al., 2022, Van Klink et al., 2024, Hartig et al., 2024]. Insect pest control, for example, uses data gathered from monitoring systems of insects that can potentially cause economic damage to decide when the best moment for controlling the popu-

lation based on a specific threshold Mitchell and Onstad [2014], Zhao et al. [2022], Van Klink et al. [2024], Parra-López et al. [2024]. In addition, managing species at risk of extinction requires a careful estimation of the targeted population's abundance [Leather, 2017, Breece et al., 2021, Randon et al., 2022]. The sustainable use of biological resources depends on monitoring the population's sustainable yield target, and examples involve the exploration of ocean fish and crustaceans Caughley [1994], Verdade et al. [2014], Krausman and Cain [2022], Alós et al. [2022], Mullowney and Baker [2023]. Observing the natural life cycle and the abundance of the target population is essential for preventing the extinction of these vital biodiversity resources [Lundgren et al., 2024].

Overall, each management action presents an opportunity for task automation, with quantitative methods tailored to address the challenges in each action. These challenges involve multiple aspects of the processes involved in each management action. For the first management action, increasing depleted populations, the use of devices in areas where species at risk of extinction have been reported, combined with algorithms that detect and classify these species, allows the increase of sample size and a more granular understanding of the status of these species. The work of Petso et al. [2021] presents an excellent example of the use of drone imaging for animal identification. Also, for the second management action, decreasing excessive populations, automation, and digitalization would also increase sample size and more granular interventions and ultimately reduce economic damage caused by pest populations, for example. For establishing maximum sustainable yields, the use of digitalization and automation allows farmers to better optimise their resources towards a more sustainable agriculture, where the sustainable yields are obtained by the solution of an optimization algorithm that takes the inputs collected from the monitoring system of agricultural processes [Kumar et al., 2025b, Shawon et al., 2023, Shetty et al., 2023, Finger, 2023]. Finally, the animal identification examples can also be applied to carry out monitoring programs without additional actions in stable populations, the sole difference being the absence of additional actions after animal identification and counting.

This thesis proposes implementing a combination of statistical, machine learning, and machine vision methods to improve the state of the art of quantitative

approaches to enhance two of Caughley's management actions: 1) carry out monitoring programs without additional actions over stable populations and 2) decrease excessive populations. The main objectives of this thesis are to apply machine vision (MV) and statistical machine learning methods to enhance the monitoring systems of animal species and to develop novel statistical learning methods to improve the state of the art in quantitative methods applied to animal control.

Recent advances in deep learning (DL) applied to classification have increased the flexibility of image classification problems [LeCun et al., 2010, Zion, 2012, LeCun et al., 2015, He et al., 2016, Mascarenhas and Agarwal, 2021, Xu et al., 2024, Gao et al., 2024, Upadhyay et al., 2025]. These improvements have found utility in solving ecological problems [Frazier and Song, 2025, Hesselbarth et al., 2025, Hu et al., 2025]. This thesis proposes an approach based on the first of Caughley's management actions to integrate DL and MV to improve animal monitoring programs. Several authors have proposed different quantitative methods for such management action presented in this thesis. Papers from 2010 until 2025 were reviewed and synthesised to show the current quantitative methods and research areas of focus related to this thesis.

Monitoring systems provide an excellent application for quantitative methods related to DL and MV [Seo et al., 2015, Liu et al., 2016, Nasir and Sassani, 2021, Sarker, 2021, Sharma et al., 2024, Balasubramaniam et al., 2025].Seo et al. [2015], Liu et al. [2016], and Nasir and Sassani [2021] present different perspectives of the importance of monitoring systems in different application highlighting the promising results of DL and MV methods and Sarker [2021], Sharma et al. [2024] and Balasubramaniam et al. [2025] present promising results of monitoring systems in smart agriculture. All of these authors present promising methods and monitoring various aspects of the agricultural system. These methods have the potential to assist entomologists in monitoring insect biodiversity [Høye et al., 2021b, Van Klink et al., 2022, Hartbauer, 2024, Leybourne et al., 2025], which is one of their many important roles [Leather, 2015, Luke et al., 2023, Collins et al., 2024]. Entomologists bring value to insect monitoring programmes by providing pathways for classifying insect species [Friedrich et al., 2014, Leather, 2015, Buckley, 2024]. For example, they associate the species with features that can be

used for selecting insect species for biological control [Luke et al., 2023, Buckley, 2024, Gerber et al., 2024]. The review articles developed by Høye et al. [2021b], Van Klink et al. [2022], Hartbauer [2024], Leather [2015], and Friedrich et al. [2014] present various perspectives on how deep learning is transforming entomology, and what entomologists' actions can benefit from the opportunities that these new methods present. Specifically, the review from Høye et al. [2021b], and Van Klink et al. [2022]. Friedrich et al. [2014] introduces other techniques more aligned with MV, such as 3D representations of insects that enable entomologists to study these specimens more effectively, and highlights the importance of entomologists' activities. Luke et al. [2023] expands on this by referring to current challenges in entomology. Moreover, Gerber et al. [2024] provides further insights into the morphology-based classification of insects, showing examples of classical methods commonly used by entomologists, and Buckley [2024] expands on this by introducing other DNA-based techniques for insect classification and highlighting the importance of monitoring insects. Leybourne et al. [2025] explores the research question "*Can artificial intelligence be integrated into pest monitoring schemes to help achieve sustainable agriculture?*" and present four criteria AI-driven systems must follow: (1) Built on accurate, efficient methods; (2) Adaptable to real-world field images; (3) User-friendly, device-based and low cost, and (4) Mobile and deployable in varied weather. These criteria reinforce the importance of developing new approaches to enhance insect monitoring systems.

The entomological classification process involves multiple steps, including the selection of anatomical and morphological features [Buckley, 2024]. In MV, pattern recognition can be used to automate insect classification [Van Klink et al., 2022, Hartbauer, 2024], and recent developments in DL applied to MV methods have demonstrated their effectiveness [Wang et al., 2024a, Li et al., 2025]. The integration of computer vision, deep learning, and machine vision methods, along with features designed based on entomologists' expertise, presents several research opportunities [Saran et al., 2025, Fotouhi et al., 2024a, Teixeira et al., 2023]. This integration can be challenging due to the differences between algorithm-based and expert-based insect classification [Schneider et al., 2023, Blair et al., 2024]. Teixeira et al. [2023] and Blair et al. [2024] explored the available deep learning methods

used for insect detection presenting the feasibility of various DL-based methods, and Schneider et al. [2023] expends the methodology search space for computer vision methods as well as mentioning the challenges of the diversity of approaches and different scales that classification can be performed, that is, the target taxa. Saran et al. [2025] explores additional data collection and methodological challenges available in insect monitoring, such as satellite imagery, unmanned aerial vehicles (UAVs), wireless sensor networks, and diversity of machine learning algorithms and in the same context, Fotouhi et al. [2024a] presents promising results of the use of DL on sticky traps. Wang et al. [2024a] and Li et al. [2025] present further examples of insect monitoring, focused on the second Caughley's management action, decreasing excessive population, demonstrating the feasibility of DL-based methods. Finally, these authors illustrate challenges of this transdisciplinary research field.

One of these challenges is to develop a method for integrating machine learning and computer vision techniques into the existing entomological framework [Fotouhi et al., 2024a, Bjerge et al., 2024, Suresh et al., 2025]. The second topic of Chapter 2 is dedicated to exploring this question. Automated classification is also challenged when identifying species with similar features, and for these cases, the role of entomologists is essential [Schneider et al., 2023, Blair et al., 2024, Nanni et al., 2025]. So, maintaining them in the classification loop during the development of new automated systems is a key component of this interdisciplinary research [Suresh et al., 2025, Cuff and Watt, 2025, Guralnick et al., 2024, Miao et al., 2021, Roy et al., 2024]. Overall, the primary objective of this field is to develop tools that assist entomologists in insect classification [Teixeira et al., 2023, Hartbauer, 2024, Rebelo et al., 2020, Cuff and Watt, 2025]. The levels of assistance range from highlighting features for one specimen in a microscope [Steinke et al., 2024, Wang et al., 2024c, Mathys et al., 2024] to automating insect abundance counts per species in the field [Ong and Høye, 2025, Ullah et al., 2024, Fotouhi et al., 2024b, Roy et al., 2024].

Another challenge is highlighted in situations where there is a restricted number of specimens of a studied species due to difficulties in collecting them in their natural habitat [Khurana, 2023, Khairunniza-Bejo et al., 2024, Fotouhi et al., 2024a]. The

lack of specimens will impact the availability of training data. The scarcity of data is a problem for deep learning methods because they require multiple observations to achieve better generalisation [Khurana, 2023, Bansal et al., 2022]. Therefore, the entomologists' perspective on this problem proves essential in identifying features that can distinguish these rare species from others [Suresh et al., 2025, Cuff and Watt, 2025, Guralnick et al., 2024, Miao et al., 2021, Roy et al., 2024]. One way to integrate these features directly into the classification pipeline is to use MV to extract these features, assisting entomologists in this scenario [Steinke et al., 2024, Wang et al., 2024c, Zhao et al., 2023]. This question is explored by the third topic of Chapter 2, where we analyse this problem using different methods to address the lack of data and class imbalance.

The stated research questions also relate to other areas of ecology that utilise animal monitoring. This thesis also focuses on the multidisciplinary research area of soundscape ecology. This area aims to understand how organisms interact with their environment by relating the acoustic features of a site to biological, geophysical, and human sound [Gasc et al., 2017, Pijanowski et al., 2024]. Ornithologists and ecologists interested in natural sounds, such as the ones produced by avian species, use several tools for passive audio recording [Molina-Mora et al., 2024, Winiarska et al., 2024]. Based on the collected data, they identify the animals encountered in the sampled region [Briseño-Jaramillo et al., 2025, Revathi and Sasikaladevi, 2025]. The identification of avian species based on their vocalisations is a time-consuming task, commonly employed by specialists who use visual representations of the specimens' vocalisations, such as spectrograms, to identify patterns specific to particular avian species in a given audio file [Potamitis, 2015, Priyadarshani et al., Gavali and Banu, 2025, Allen-Ankins et al., 2025]. One of the time-consuming tasks in this research is identifying audio files that do not contain any vocalisations of interest, as they may be more abundant than those with the presence of the species of interest [Zhang et al., 2013, Aide et al., 2013, Gibb et al., 2019, Kershenbaum et al., 2025]. Thus, the combination of deep learning and machine vision can benefit these researchers by eliminating the need for these time-consuming tasks [Maithripala et al., 2024, Das et al., 2024, Kumar et al., 2025a].

Several authors have proposed DL and MV methods for automating the passive monitoring of avian species based on audio trackers [Kahl et al., 2021, Liu et al., 2022, Sharma et al., 2023, Wei et al., 2024, Le Penru et al., 2025, Heinrich et al., 2025]. Recurrent neural networks and classical machine learning methods, such as support vector machines, have been proposed to identify avian species based on audio frequencies [Srujana et al., 2023, Farah et al., 2024, Abhinav and Dhavale, 2024]. Over the past few years, significant results have highlighted the high accuracy of Convolutional Neural Networks (CNNs) in classifying avian species based on spectrogram images and other data sources [Suryavanshi et al., 2024, ME et al., 2024, Adhikari et al., 2024, Arthy et al., 2025]. However, considering the vast diversity of avian species, classification performance varies in application, which inspired the combination of hybrid approaches that combine classical feature engineering methods with the use of transfer learning or solely a CNN architecture trained from scratch targetting better performances for more challenges cases [Lostanlen et al., 2019, Pahuja and Kumar, 2021, Gupta et al., 2021, Mohaghegh et al., 2023, Wang et al., 2024b, Segura-Garcia et al., 2024]. The first topic of Chapter 2 explores the combination of other traditional computer vision techniques to preprocess spectrograms with transfer learning based on a classical CNNs architecture.

The second of Caughley's management actions targeted by this thesis relates to decreasing excessive populations. This action is commonly encountered in the daily activity of agricultural research when studying the best approach for controlling insect pests, for example [Tajmiri et al., 2017, Zainudin et al., 2023, Rajabpour, 2025]. Integrated Pest Management (IPM) is a research field that aims to develop approaches for controlling insect pests by reducing their population density [Barzman et al., 2015, Deguine et al., 2021b]. The primary challenge in this field is determining the optimal moment to apply a method that reduces population density, such as biological, chemical, or physical control [Benseddik et al., 2022, Byrne et al., 2025]. Also, a combination of these approaches can be selected. However, the question of when is the right moment to apply a method to reduce population density remains open [Reilly and Elderd, 2014, Karmawati et al., 2025]. Authors have proposed using different strategies based on insect population density levels, such as establishing an economic threshold [Ragsdale et al., 2007, He et al., 2020]

to predict the optimum intervention moment.

The application of statistical learning methods yielded promising results, providing valuable insights to address the stated problem Kishi et al. [2023b], Singh et al. [2024b], Kumari et al. [2025]. The combination of biological features of the studied population provides information for training statistical learning algorithms. Potential features may be patterns of population density that occur prior to an event of interest [Burant et al., 2021, Kiobia et al., 2025]. When the population density of a given insect pest species exceeds the accepted threshold, it indicates an outbreak event [Paredes et al., 2022]. The definition of a threshold depends on the specific insect pest population being studied [Bueno et al., 2010, Chasen et al., 2015, Jiménez et al., 2021]. Also, specialists consider the potential economic damage caused by the insect pest to determine this threshold [Chasen et al., 2015, Zhou et al., 2024, Reisig and Huseth, 2025]. However, even with a defined threshold, there is still a need to predict these events, mainly extreme ones [Djaman et al., 2019, Lawton et al., 2022, Tajdar et al., 2025]. The combination of statistical learning and population patterns has the potential to develop a novel interpretable statistical learning method. This question is explored by the first topic of Chapter 3.

Insect pest population dynamics depend on several biological, chemical and physical properties of the studied pest species and their surrounding environment [Lawton et al., 2022, Boulanger et al., 2025]. Authors have demonstrated the influence of climatic variables on the development of multiple insect pest species [Huang and Li, 2015, Huang and Hao, 2020, Bapatla et al., 2022]. The studies typically employ controlled experiments to confirm such a causal influence Tajdar et al. [2025]. However, capturing the causal influence of climate variables on insect populations in monitoring systems is more challenging than experiments due to random variation impacting population densities Tsoumas et al. [2025], Saberski et al. [2024]. Therefore, there is a need to develop new methods that aid researchers in investigating the causal relationship between these climatic variables and the insect population, thereby improving insect abundance forecasting [Runge, 2023, Zipkin and Doser, 2024]. The development of causal models is commonly encountered in environmental science, climate change and other fields [Tsoumas et al., 2023]. The

combination of causal analysis with the forecasting potential of statistical learning algorithms presents an opportunity for further research and methodological development, particularly in the entomological context. The second topic of Chapter 3 explores this question by developing a novel framework for insect forecasting that combines statistical learning, dynamic systems techniques and Granger's causality.

Overall, in chapter 2, we address the following research questions directly related to conservation biology by examining Caughley's management actions. In the first application of chapter 2, this thesis explores the research question "Can spectrogram improve audio-based classification of nocturnal Brazilian avian species?" based on the hypothesis that combining CNNs with this signal representation can achieve accurate classification. For the second application, this thesis explores the research question "Can machine learning and computer vision methods be integrated into the existing entomological framework?", with the independent hypothesis that integrating concepts of classical machine vision within the entomological context is feasible. Finally, for the third application, we explore the following research question: "What combination of ML methods, new features and augmentation techniques can enhance Anastrephas's species classification?" and the hypothesis associated with this question is based on the hypothesis that exploring various agumentation and machine learning techniques using a new wing's structure features can be a potential search space that allows the finding of better approaches. For chapter 3, explore the following research questions directly associated with conservation biology by exploring Caughley's management actions. In the first application, we explore the question "Can numeric patterns before an outbreak event be used to predict this phenomenon?" and the hypothesis associated with the question is that designing a new machine learning method based on a similarity structure for patterns preceding an outbreak can provide competitive performance and be a transparent method to practitioners. Finally, the second application explores the question "Can statistical learning algorithms combine with causal analysis improve insect abundance forecasting?" and hypothesizes that the combination of statistical learning and dynamical systems with Granger's causality can be competitive with the current state of the art.

1.2 Datasets used in this thesis

To explore the questions stated by this thesis, in the first topic of Chapter 2, we used three datasets, with respectively histogram-equalised, grayscale, and coloured spectrograms. All datasets contain 505 spectrogram images of size 150×150 with 6 classes containing 60-second audio clips of (1) *Antrostomus rufus*'s vocalisation, (2) *Megascops choliba*'s vocalisation, (3) both species vocalising together, (4) *Antrostomus rufus* then *Megascops choliba* vocalisation, (5) *Megascops choliba* then *Antrostomus rufus* vocalisation, and (6) background only sounds.

In the second topic of Chapter 2, we used a dataset of medically and forensically important flies [Ong and Ahmad, 2022] to illustrate the machine vision pipeline into the existing entomological classification framework. The dataset contains a set of classified images which includes the genus *Chrysomya* (731), *Lucilia* (587), *Sarcophaga* (570), *Rhiniinae* (488), and *Stomorhina* (500).

In the third topic of Chapter 2, a novel dataset of *Anastrepha* species from the *pseudoparallela* group was used. The dataset will help elucidate the research questions related to assisting entomologists with classification problems for rare species. The dataset contains images of the right-wing of 127 females from five species of this group.

For both topics of Chapter 3, we used data from an aphid monitoring programme implemented in Southern Brazil (State of Rio Grande do Sul, RS) from 2011 to 2019, totalling 424 observations. The dataset contains a weekly total of collected aphids and climate covariates, including temperature, humidity, wind speed, and other relevant features.

1.3 Outline of the thesis

The remainder of this thesis is organised as follows. In the first topic of Chapter 2, we explored pre-processing techniques necessary for training CNNs applied to avian species classification. We proposed using a pre-trained VGG16 CNN architecture to identify two nocturnal avian species, namely *Antrostomus rufus* and *Megascops choliba*, commonly encountered in Brazilian forests. Our collaborators recorded sounds in 16-bit wave files at a sampling rate 44Hz and classified the

presence of these species. With the classified wave files, we created additional classes to visualise the performance of the VGG16 CNN architecture for detecting both species. We had six categories containing 60 seconds of audio of species vocalisation combinations and background-only sounds. We produced spectrograms using the information from each RGB channel, only one channel (grey-scale), and applied the histogram equalisation technique to the grey-scale images. A comparison of system performance using histogram-equalised images and unmodified images. Investigating the effect of the image pre-processing on the performance of the CNN was a feature of this study. Moreover, to show the practical application of our work, we created 51 minutes of audio, which contains more noise than the presence of both species (a scenario commonly encountered in field surveys).

The second topic of Chapter 2 presented a proposal for integrating machine learning and computer vision methods into the existing entomological framework. The chapter introduced applications of machine vision methods to identify and localise insects and features of their anatomy. We introduced the machine vision pipeline, image descriptors and fundamental methods, such as thresholding, blob and contour detection. Moreover, deep learning methods are applied for insect classification, including deep neural networks, convolutional neural networks and concepts of transfer learning. We presented a U-Net architecture trained with the dataset of medically and forensically important flies for insect localisation. Finally, we applied the dimensional reduction method PaCMAP to visualise the features extracted from the fly datasets. The Grad-CAM method is used on the flies datasets to assist insect object localisation.

The third topic of Chapter 2 contributed to automating the classification of the species *Anastrepha consobrina*, *Anastrepha curitis*, *Anastrepha dissimilis* and *Anastrepha sp colombo* based on DL methods. This species group is known as pseudoparallela, and specimens of these species are economically important pests of fruit crops. We used a three-fold cross-validation for training, tuning, and testing, encompassing six permutations of these steps to assess the performance of the learning algorithms. We explored transfer learning solutions and proposed a new set of features based on each wing's structure. We used the dual annealing algorithm to optimise the hyperparameters of Deep Neural Networks, Random Forests,

Decision Trees, and Support Vector Machines algorithms. Given the scarcity of data, we utilised autoencoders and SMOTE algorithms to address the class imbalance and augment the number of data samples from each specimen in the original dataset. The utilisation of data augmentation methods for this context and the effect on classification was discussed in the chapter.

In the first topic of Chapter 3, we propose the Pattern-Based Prediction (PBP) method for predicting population outbreaks. It uses information on previous time series values that precede an outbreak event as predictors of future outbreaks, which can be helpful when monitoring pest species. We illustrate the methodology using simulated datasets and an aphid time series obtained in wheat crops in Southern Brazil. We benchmarked our results against established state-of-the-art machine learning methods: Support Vector Machines, Deep Neural Networks, Long Short Term Memory and Random Forests. The implemented PBP method is available in Python through the `pypbp` package (https://pypbp-documentation.readthedocs.io).

In the second topic of Chapter 3, we propose a new approach to select the appropriate time for applying an intervention for reducing the insect pests based on the causal effect of climate covariates and their abundance. We address this problem by combining statistics, machine learning, and time series embedding. We pre-processed the data using our newly proposed approach and more straightforward approaches. We used a Random Forests algorithm to show that our novel approach yields competitive forecasts as indicated by the Root Mean Squared Error obtained from test data. Finally, we compared the RF performance based on one-step-ahead forecasts using the original dataset with all features, the dataset obtained from the proposed causal approach, and two datasets based on insect abundance, with delays of 3 and 6 time steps.

All proposed methods in this thesis were implemented using the R [R Core Team, 2025] and Python software. They are accessible on the author's Github[1] via three public repositories. Finally, in Chapter 4, we conclude the thesis by proposing topics for future research.

[1] https://github.com/GabrielRPalma

CHAPTER 2

Machine Vision applied to animal monitoring systems

This chapter's contributions to Caughley's management action related to monitoring programs without additional actions over stable populations are based on the analysis of preprocessing and feature engineering techniques applied to image and audio-based monitoring of animals. The main focus of this Chapter is to propose new approaches combining machine vision and learning methods to classify and localise animal species. This Chapter explores the applications of feature engineering based on transfer learning and computer vision to aid animal monitoring by introducing approaches for dealing with challenges in animal classification, such as small and imbalanced datasets.

Overall, the Chapter will review current literature focused on the automatic classification of avian and insect species as case studies to present applications of the combination of the discussed techniques. Three applications will be presented in this Chapter. The first one describes the identification of avian species in the context of soundscape ecology. The second application involves entomology, specifically medically and forensically important flies of the order Diptera. The last application also involves entomological applications. However, the focus is on fruit fly species of the *Anastrepha pseudoparallela* group (Diptera: Tephritidae)

[Araújo et al., 2024].

2.1 A machine vision system for avian song classification with CNN's

The contribution of this thesis application to the target Caughley's management action focuses on the context of soundscape ecology, specifically on using spectrograms of avian vocalisations as input data for pre-trained CNN classification. The demonstrated feasibility of pre-trained CNNs has direct implications for future work on audio-based classification of avian species, given the challenges of manual identification. The manual identification process involves using sound-based features, such as the frequency content of a given signal. Also, specialists studying avian species can further convert this signal into spectrograms to manually determine their scientific names.

Based on the promising results of deep learning methods, such as Convolution Neural Networks (CNNs) in image classification, here we propose the use of a pre-trained VGG16 CNN architecture to identify two nocturnal avian species, namely *Antrostomus rufus* and *Megascops choliba*, commonly encountered in Brazilian forests. Monitoring the abundance of these species is important to ecologists to develop conservation programmes, detect environmental disturbances and assess the impact of human action. Specialists recorded sounds in 16-bit wave files at a sampling rate of 44kHz and classified the presence of these species. With the classified wave files, we created additional classes to visualise the performance of the VGG16 CNN architecture for detecting both species. We end up with six categories containing 60 seconds of audio of species vocalisation combinations and background only sounds. We produced spectrograms using the information from each RGB channel, only one channel (grey-scale), and applied the histogram equalisation technique to the grey-scale images. A comparison of the system performance using histogram equalised images and unmodified images was made. Histogram equalisation improves the contrast, and so the visibility to the human observer. Investigating the effect of histogram equalisation on the performance of the CNN was a feature of this study. Moreover, to show the practical applica-

tion of our work, we created 51 minutes of audio, which contains more noise than the presence of both species (a scenario commonly encountered in field surveys). Our results showed that the trained VGG16 CNN produced, after 8000 epochs, a training accuracy of 100% for the three approaches. The test accuracy was 80.64%, 75.26%, and 67.74% for the RGB, grey-scaled, and histogram equalised approaches. The method's accuracy on the synthetic audio file of 51 minutes was 92.15%. This accuracy level reveals the potential of CNN architectures in automating species detection and identification by sound using passive monitoring. Our results suggest that using coloured images to represent the spectrogram better generalises the classification than grey-scale and histogram equalised images. This study might develop future avian monitoring programmes based on passive sound recording, which significantly enhances sampling size without increasing cost.

2.1.1 Introduction

As described in session 1, soundscape ecology involves studies of bioacoustics, landscape ecology, community ecology and engineering to answer research questions. The combination of those disciplines benefits the zoological and ethological communities by providing a theoretical framework grounded in a broad ecological context, a wealth of long-term soundscape collections from around the world, methods optimising acoustic monitoring, and the analysis of acoustic big-data Gasc et al. [2017].

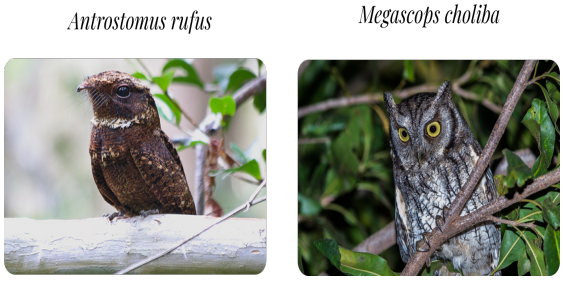

Figure 2.1: *Antrostomus rufus* and *Megascops choliba* species. These pictures were provided respectively by Rafael Cerqueira and Rafael Martos Martins.

The collection of acoustic data involves a high volume, high acquisition rate and a variety of signals Gasc et al. [2017], Pijanowski et al. [2011], showing the necessity of new methods to analyse acoustic big-data
Emmanuel and Stanier [2016], Brunsdon and Comber [2020]. Consequently, the

zoological and ethological communities can also benefit from Deep Learning (DL) techniques.

Creating new methods that automate identifying avian species based on spectrograms will provide several gains for the zoological and ethological communities. Among the possible ways to analyse signals from natural sounds, the Fourier transformation coupled with the Gabor transformation offers many benefits for a visual interpretation of avian songs. However, understanding spectrograms is a challenge for biologists requiring effort to train researchers to detect species. This necessity of rigorous training to achieve high accuracy in classifying avians led to an opportunity for the application of deep learning methods Zhang et al. [2019], Ruff et al. [2020, 2021], Hidayat et al. [2021], Permana et al. [2021], Bravo Sanchez et al. [2021]. Brazilian biomes represent excellent opportunities to study the application of these methods given the vast diversity of avian species. Therefore, this chapter has two main objectives: i) to implement an algorithm capable of detecting two Brazilian species of nocturnal avians, and ii) to evaluate the impact of image pre-processing on species classification.

2.1.2 State of the Art

Spectrograms have been widely used by specialists seeking to identify avian species by their vocalisations [Jahn et al., 2017]. As this manual task of inspecting spectrogram and identifying features can be time-consuming, especially based on the amount of audio files with the absence of relevant signals, several researchers have implemented Deep Learning (DL) algorithms to automate this process [Gupta et al., 2021, Garcia et al., 2024, Heinrich et al., 2025]. Also, DL has been applied in soundscape ecology, zoology and ethology research projects were primarily interested in species identification Selin et al. [2006], Chou et al. [2007], Sprengel et al. [2016], Lasseck [2018b], Christin et al. [2018], Sankupellay and Konovalov [2018], Lasseck [2018a], Zhang et al. [2019], Koh et al. [2019], Ruff et al. [2020], LeBien et al. [2020], Ruff et al. [2021], Huang and Basanta [2021], Campos Paula et al. [2022]. Widely used algorithms in this context are Deep Neural Networks and Convolutional Neural Networks (CNNs) Ruff et al. [2020], Christin et al. [2018], Zhang et al. [2019], Ruff et al. [2021], Hidayat et al. [2021], Kahl et al. [2021],

Permana et al. [2021], Disabato et al. [2021]. Over the last year, studies showed good Deep Learning performances on avian species identification based on their sounds using mainly the Deep Neural Networks and CNN architectures Kahl et al. [2021], Ruff et al. [2021], Hidayat et al. [2021], Permana et al. [2021]. These researchers focussed their effort on classifying species and applying different image pre-processing techniques. An example of this approach was evaluating the effect of grey-scale and jet colour map Spectrogram on the accuracy of avian species classification Incze et al. [2018].

2.1.3 Methods

We experimented with 20 autonomous acoustic recording units in Angatuba (São Paulo - Brazil) to monitor two nocturnal species. Specialists recorded sounds in 16-bit wave files at a sampling rate of 44kHz. These recordings followed a discontinuous protocol (1-minute recordings in 3-minute intervals) 24 hours a day, for 15 days during 6 months.

We selected nocturnal avians because, at night, other sounds such as biophony, and anthrophony are not so evident Gasc et al. [2017]. Also, these species are common in Brazil and produce sounds with almost no variation and constant amplitude. With the classified wave files, we created additional classes to visualise the performance of the VGG16 CNN method (see Figure 2.3 for details) for detecting both species. We created 6 classes containing 60-second audio clips of (1) *Antrostomus rufus*'s vocalisation, (2) *Megascops choliba*'s vocalisation, (3) both species vocalising together, (4) *Antrostomus rufus* then *Megascops choliba* vocalisation, (5) *Megascops choliba* then *Antrostomus rufus* vocalisation, and (6) background only sounds. Specialists classified the presence of *Antrostomus rufus*, *Megascops choliba*, and the background sound using the Raven Pro 1.4 software (Bioacoustic Research Program, 2011). Experts in this field performed the provided classification, and labels were assigned to every 60-second audio segment based on features commonly used to identify these species.

Then, based on the Gabor transformation of the audio frequency data, using the spectrogram we produced colour images - one per RGB channel, only one channel (grey-scale), and applied the histogram equalisation technique on the grey-scale

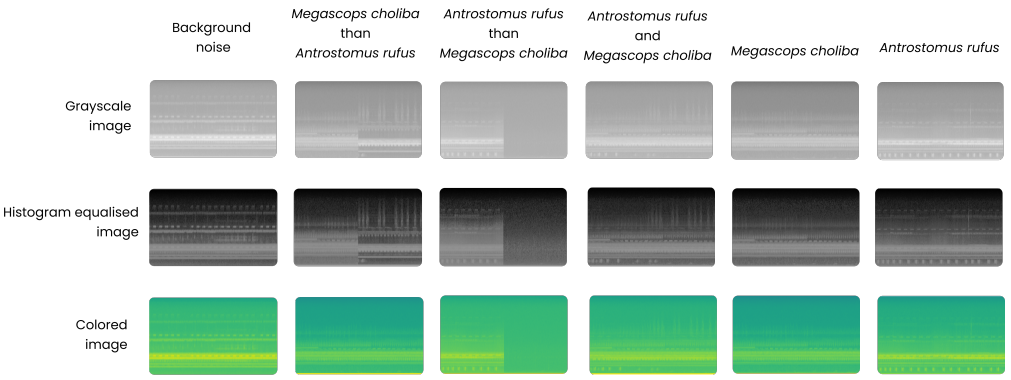

Figure 2.2: spectrograms of *Antrostomus rufus*'s vocalisation, *Megascops choliba*'s vocalisation, both species vocalising together, *Antrostomus rufus* than *Megascops choliba* vacalization, *Megascops choliba* than *Antrostomus rufus* vocalisation. The diagram also shows the differences between grey-scale, histogram equalised and coloured images.

images to better visualise the contrasts of the spectrograms. Histogram equalisation was used to maintain the spatial and dynamic range properties of the image. Other enhancements, such as Gaussian filtering and thresholding were rejected in this study as they reduce the information of the image. Thus, we compiled a data set containing 505 images of size 150×150 (Figure 2.2), the dataset contained 98, 43, 100, 100, 100, and 65 images for respectively (1) *Antrostomus rufus*'s vocalisation, (2) *Megascops choliba*'s vocalisation, (3) both species vocalising together, (4) *Antrostomus rufus* then *Megascops choliba* vocalisation, (5) *Megascops choliba* then *Antrostomus rufus* vocalisation, and (6) background only sounds. The feature extraction was based on convolution and pooling operations, and the parameters were optimised using the ImageNet data set (Chollet [2018]; Wani et al. [2020]; Simonyan and Zisserman [2014]). Then, using the feature maps with size 4×4 provided by the pre-trained model, we trained two additional densely connected layers with a dropout rate of 0.5 to identify the nocturnal avian species (see Figure 2.3 for an illustration of these operations). We used 80% of the data (412 images) for training and 20% (93 images) to test the proposed CNN architecture's performance. We used 8000 epochs to compute the method's accuracy, precision, and recall.

Finally, to show the practical application of our work, we created two audio files of 51 minutes, which contains more noise than the presence of both species. This is a common scenario encountered in field surveys, and it consumes a representative amount of time from researchers that analyse such data. The first audio file contains both species singing together more frequently, and in the second one, both species sing alone more regularly. To construct them, we selected spectrograms from the test data and reordered them to create the 51-minute audio file.

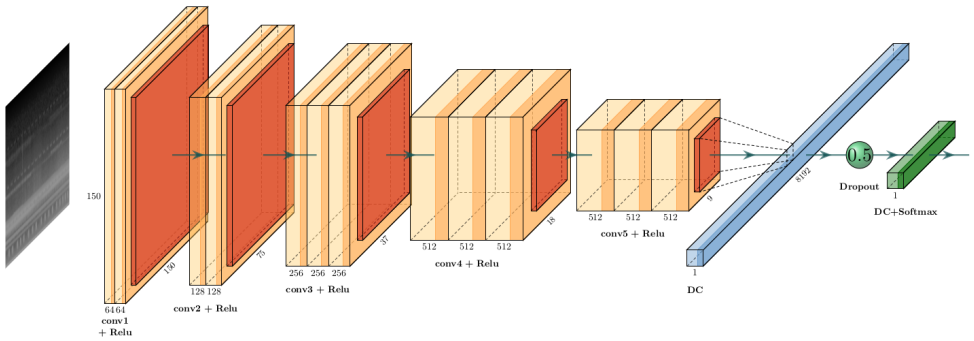

Figure 2.3: Scheme representing the pretrained VGG16 CNN architecture and the additional layers used to train the model. The convolution, max-pooling and dropout operations are represented in orange, red and green, respectively. Finally, densely connected (DC) layers are added with two activation functions: ReLU before the dropout and softmax afterwards.

2.1.4 Results

Figure 2.4 shows the accuracy, precision and recall metrics obtained from the CNN VGG16 architecture using the selected pre-processing spectrograms. The trained VGG16 CNN produced, after 8000 epochs, a training accuracy of 100% for the three approaches. The test accuracy was 80.64%, 75.26%, and 67.74% for the RGB, grey-scaled, and histogram equalised approaches. It indicates that the CNN trained with coloured images provided a better generalisation than the one based on histogram equalised or grey-scale images. The decision to explore spectrogram formats is based on the origin of the pretrained dataset of the proposed CNN. As VGG16 was pretrained on RGB images, one possible explanation for the better accuracy may lie in this. However, we found this result curious, and further investigation is needed to reach concrete conclusions.

Considering the experimental optimised parameters of the CNN introduced with coloured pictures, the accuracy on the synthetic audio file of 51 minutes presented in Figure 2.5 was 92.15% and in Figure 2.6 was 76.47%. The sequence created for Figure 2.5 investigated the system's performance when the species are vocalising at the same time. Figure 2.6 focusses on the performance when the species sing at different times. It indicates that the system with this approach would help the researcher identify the presence of these two species on a large dataset containing more noise than the individuals themselves. In Table 2.1, we present the confusion matrix provided by the VGG16 architecture trained utilising the information from each RGB channel.

Finally, this pilot study shows promising results indicating an optimistic scenario for improved DL studies of this kind to automate the detection of nocturnal avian species in Brazil. We identify the following limitations of this study. The first relates to the class imbalance, which was not accounted for in the analysis. Secondly, we have data scarcity in both the training and testing sets. The study demonstrates the feasibility of using spectrograms to classify the studied avian species; however, further work with larger sample sizes is needed to improve model generalisation. That work should implement early stopping rules, show learning curve diagnostics, and include a baseline to make comparative performance judgments. The final limitation concerns the limited number of nocturnal species studied. Our study demonstrates the feasibility of classifying these species in a highly restricted scenario, where little noise is present in the audio files. Therefore, further investigation of different periods, locations, and even seasons provides fruitful future work.

2.1.5 Conclusion

The obtained validation accuracy shows the feasibility of the pre-trained VGG16 architecture in detecting the studied avian species. Also, given the number of classes presented in this this chapter, including the presence of both species and noise, our results show a good perspective for further investigation of soundscape studies, including other species. Our results suggest that using coloured images to represent the spectrogram generalises the classification better than grey-scale and

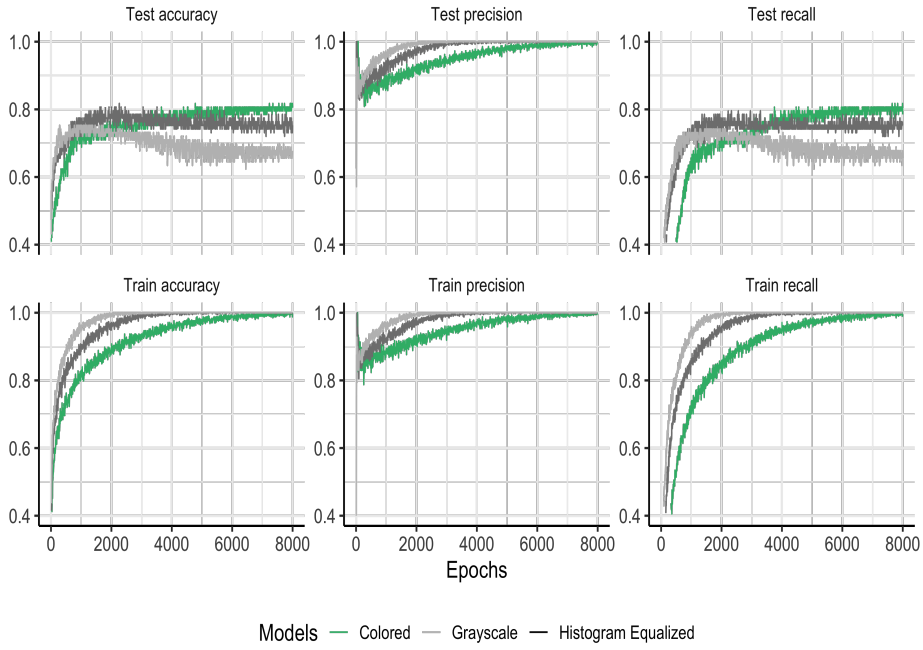

Figure 2.4: Accuracy, precision and recall metrics of the CNN VGG16 architecture for the train and test data using 8000 epochs. Green, grey and black lines result from architectures trained utilising the information from each RGB channel, only one channel (grey-scale), and applied the histogram equalisation technique on the grey-scale images.

histogram equalised images. This study will serve as a basis for developing a future animal monitoring program based on passive recording sound, which significantly enhances sampling efforts without increasing cost. Finally, as future work, we will explore new advancements in end-to-end models that process audio as a time series using architectures, such as transformers.

Predicted	Observed					
	Antrostomus rufus before *Megascops choliba*	*Antrostomus rufus*	Both species	*Megascops choliba* before *Antrostomus rufus*	*Megascops choliba*	Noise
Antrostomus rufus before *Antrostomus rufus*	19	0	0	1	0	0
Antrostomus rufus	0	11	5	0	1	0
Both species	0	3	13	0	3	0
Megascops choliba before *Antrostomus rufus*	1	0	0	17	0	0
Megascops choliba	0	0	3	0	4	0
Noise	0	0	1	0	0	11

Table 2.1: A confusion matrix produced based on the predictions of the CNN VGG16 architecture trained utilising the information from each RGB channel.

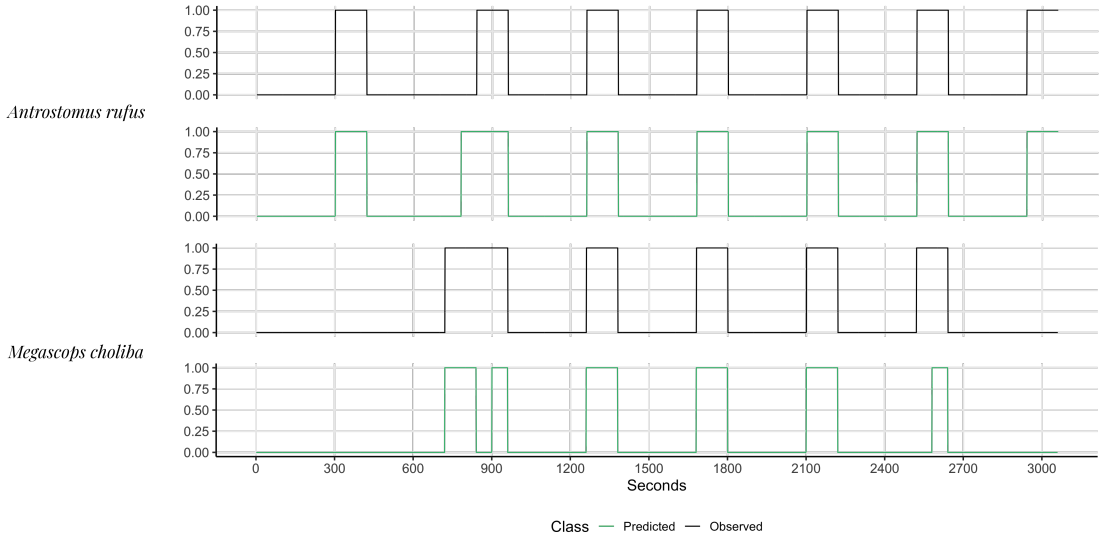

Figure 2.5: Detection of the studied species based on the CNN VGG16 architecture using 51 minutes of audio showing the practical application of our results. The green line represents the detection of the CNN and the black line is the real class detected by the specialists. For this dataset, we obtain an accuracy of 92.15%.

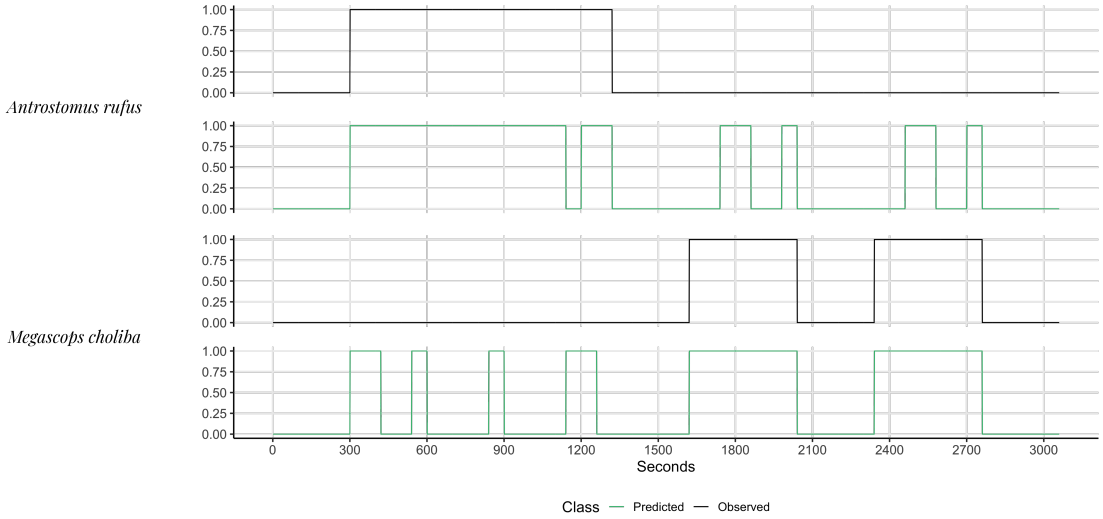

Figure 2.6: Detection of the studied species based on the CNN VGG16 architecture using 51 minutes of audio showing the practical application of our results. The green line represents the detection of the CNN and the black line is the real class detected by the specialists. For this dataset, we obtain an accuracy of 76.47%.

2.2 Machine Vision Applied to Entomology

The contribution of this next thesis application to the target Caughley's management action presents a proposal for integrating machine learning and computer vision methods into the existing entomological framework. The application cover existing techniques involving Machine vision to solve practical challenges faced by the entomological community. Machine vision is an interdisciplinary field that combines methods from computer vision and machine learning, mainly deep learning to solve issues. Common issues, such as classification and localisation are typical examples that combine these research fields. These techniques have many applications in entomology. In the entomological context, classification methods identify insect species or features of an insect, and localisation methods return their position within a scene. The available techniques for localisation and classification benefit the entomological community by assisting specialists. This application introduces the use of machine vision methods to identify and localise insects, and the features of their anatomy. The machine vision pipeline, image descriptors and fundamental methods, such as thresholding, blob and contour detection, are presented. Deep Learning (DL) methods are applied for insect classification,

focusing on Deep Neural Networks and Convolutional Neural Networks (CNNs), including concepts of transfer learning applied to insect classification. A dataset of medically and forensically important flies, Diptera are used to illustrate these methods. CNN-based methods are described for insect localisation, and semantic segmentation using the U-Net architecture trained on the dataset of flies is presented. The dimensional reduction method PaCMAP is applied to visualise the feature extracted from the fly datasets. The Grad-CAM method is used on the flies datasets to assist insect object localisation. Finally, different platforms of computing are presented and compared under three different metrics of performance, power consumption and operating temperature.

2.2.1 Introduction

Taxonomic classification is one of the pillars required for implementing Integrated Pest Management (IPM) methods Jaleel et al. [2018], Deguine et al. [2021a], Shimbori1 et al. [2023]. It allows the identification of important pests with a detailed description of their biology. Researchers use the biological information gathered by entomologists to create management theories, yielding essential technologies to control the insect pest population Stenberg et al. [2021]. These new findings positively impact our society, for example, the economic gains that biological control has offered to agriculture. These gains can reduce environmental impact and cost and enhance food security Bale et al. [2008], van Wilgen et al. [2020], Stenberg et al. [2021].

Entomological research has provided direct contributions to the control of pest management in agriculture, such as the introduction of the species *Tamarixia radiata* that allowed the biological control of the pest *Diaphorina citri* Étienne et al. [2001], Qureshi et al. [2009], Chow and Sétamou [2022]. The description and identification of parasitoid species *Trichogramma galloi* (Zucchi, 1988) acted as an important insect for the management of *Diatraea saccharalis* Cônsoli and Parra [1996], Camarozano et al. [2021]. These studies also present the spatial distribution of the targeted species and recommendations about the parasitism rate. The

biological descriptions allow further research and quantitative approaches, such as mathematical, statistical and computational modelling Camarozano et al. [2021].

In addition, entomological research offers vital information to wildlife conservation by gathering awareness of global biodiversity and providing evidence of collected species with clear documentation in national museums as evidence. As this is such an important field, many researchers have tried to develop techniques to automate insect identification. By combining the collected knowledge from computer vision, deep learning, and labelled datasets created by entomologists, it is now possible to create algorithms that can automate or assist the classification of insects in different taxonomical levels, such as order Ozdemir and Kunduraci [2022], genus Ong and Ahmad [2022] and species Thenmozhi and Reddy [2019].

Several researchers have implemented computer vision techniques to automate insect identification of species such as *Bemisia tabaci*, *Sesamia inferens*, *Chilo suppressalis*, Zayas and Flinn [1998], Zhigang et al. [2003], Larios et al. [2008], Solis-Sánchez et al. [2009], Yang et al. [2010], Asefpour Vakilian and Massah [2013], Favret and Sieracki [2016], Qing et al. [2020], Kasinathan and Uyyala [2021], Kasinathan et al. [2021]. These results have shown promising results by obtaining features and applying machine learning algorithms, such as Support Vector Machines (SVM), Deep Neural Networks (DNNs), Linear Discriminant Analysis (LDA), and Random Forest (RF) Hastie et al. [2004], Goodfellow et al. [2016], Mello and Ponti [2018], Wani et al. [2020] to classify the species. These results illustrate the feasibility of combining machine learning and vision methods.

The introduction of deep Learning methods based on Deep Neural Networks has expanded machine vision applications and enhanced the performance of the available methods. With this new toolkit of methods, the identification and localisation of insects within images entered a new stage of accuracy. Several papers presented the applications of deep learning methods on insect identification and localisation of *Apis melifera*, *Spodoptera frugiperda*, *Diabrotica speciosa*, fruit flies of the genus *Anastrepha*, and different species of mosquitoes including *Aedes aegypti* Leonardo et al. [2018a], Dawei et al. [2019], Peng et al. [2019], Patel and Bhatt [2021], Naufal et al. [2021], Shen et al. [2021], Martins et al. [2019a], Souza et al. [2019], Tetila

et al. [2019], Hansen et al. [2020], Kaur et al. [2022], Silveira et al. [2021], Zhang et al. [2022], Feng et al. [2022]. These examples showed promising results for the implementation of automated insect identification based on the combination of machine vision and deep learning methods.

Multiple methods based on Deep Neural Networks (DNN) have been used to identify insect species based on extracted features from the insects within images Huynh et al. [2019], Shi et al. [2020], Toscano-Miranda et al. [2022], Salifu et al. [2022]. Feature extraction consists of using machine vision, signal processing and statistical methods or a combination of these methods to obtain relevant features of an insect that are then used to identify and train DNNs or other machine learning techniques to classify the selected insect species Szeliski [2022]. Principal Component Analysis (PCA), sparse coding, and scale-invariant feature transforms are examples of these methods Martineau et al. [2017], Kasinathan and Uyyala [2021], Kasinathan et al. [2021]. Biological features can be extracted based on geometric morphometrics techniques, such as the Elliptic Fourier features method Tatsuta et al. [2018]. Geometric morphometrics involves techniques to study scale and shape relationships of structures using Cartesian geometric coordinates rather than linear, areal (of area), or volumetric variables.

One of the essential techniques that linked machine vision with deep learning methods was the introduction of Convolutional Neural Networks (CNNs) LeCun et al. [1989]. CNN-based methods include a variety of architectures that provide higher performance for classification and localisation tasks Goodfellow et al. [2016], Wani et al. [2020]. These architectures vary in the number and type of operations used in their pipeline. Several papers used different CNN architectures to automate the identification of multiple taxa of insects, including VGG-16, LeNet, AlexNet, Xception and EfficientNet architectures. Novel methods combining CNN methods with metadata have shown promising results for insect identification Thenmozhi and Reddy [2019], Tetila et al. [2019], Kuzuhara et al. [2020], Wang et al. [2021], Høye et al. [2021a], Rimal et al. [2022], Wang [2022]. Additionally, for localisation tasks, methods such as U-Net Ronneberger et al. [2015b], R-CNN Girshick et al. [2014], Fast R-CNN, Cascade R-CNN, and You Only Look Once (YOLO) are commonly used in research Høye et al. [2021a], Zha et al. [2021], Cheng et al.

[2022], Hong et al. [2022].

The main goal of this chapter is to introduce the application of vision methods used to identify and localise insects. In Section 2.2.2, the machine vision pipeline, image descriptors and fundamental methods, such as thresholding, blob and contour detection are presented. In section 2.2.3, Deep Learning (DL) methods are applied for insect classification, focusing on Deep Neural Networks and Convolutional Neural Networks (CNNs), including concepts of transfer learning applied to insect classification. To illustrate these methods, a dataset of medically and forensically important flies *Diptera* are used to illustrate these methods. In Section 2.2.4, CNN-based methods are described for insect localisation, and semantic segmentation using the U-Net architecture trained on the dataset of flies is presented. In addition, an object detection algorithm called You Only Look Once (YOLO) is explored. In Section 2.2.5, different computing platforms are presented and compared under three different metrics of performance, power consumption and operating temperature.

2.2.2 Machine Vision Pipeline

Computer vision focuses on the extraction of information from images. Machine vision extends the focus to the integration and automation of the techniques, converting an image into useful information that can be obtained using a standard workflow. A typical workflow that can be applied to many image processing methods is shown in Figure 2.7. Modern machine vision methods integrate machine learning algorithms into the workflow. Typically, machine learning algorithms assist with segmentation, classification and localisation.

Capture: This involves choosing and configuring suitable equipment to capture images of sufficient quality to pass into the machine vision pipeline. The main imaging devices and cameras used to capture images relating to pest species detection include microscopes, camera traps, mobile phones and SLR cameras. Other technologies, such as sonification, can also produce images Palma et al. [2022].

Specifying a camera sensor suitable for a specific application requires consideration of colour sensitivity (how close to an actual colour a camera can record), spectral

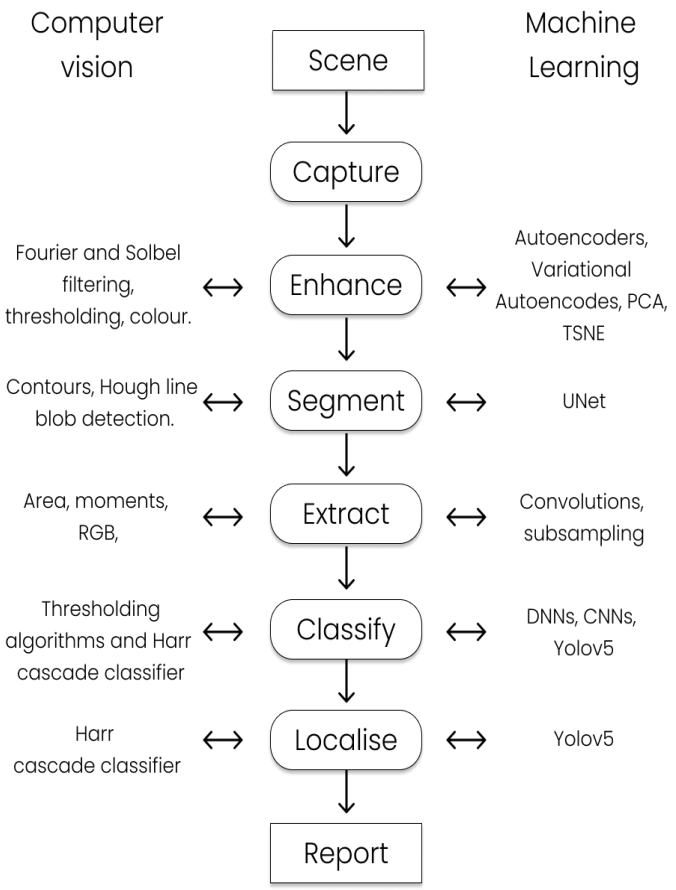

Figure 2.7: Overview of a machine vision pipeline workflow.

range (range of wavelengths a camera can capture), resolution (number of pixels in an image), dynamic range (contrast ratio between the darkest and brightest colour tones), shutter control, and framerate (number of images a camera can capture in a second to produce a video). The choice of lenses attached to the camera requires consideration of the aperture (which governs the amount of light the lens gathers), the field of view (the extent of the world visible to the camera), distortion (the aberrations that can cause straight lines to curve), and chromatic aberration (the effects caused by lens focus varying with colour). Lenses have standard mounts and are designed for specific sensor sizes. Adjustable lenses can change the field of view, apperture and adjust the focus. Lighting is an important component of image capture. The correct choice of lighting can reduce shadow and prevent

motion blurring. Multispectral and coloured light sources can be used to enhance features in the scene.

A pinhole camera is the fundamental computational model used in computer vision. Real cameras use lenses which introduce distortion. In addition, the lenses can be misaligned with the imaging sensor that detects information to make an image. In most cases, a camera is calibrated using multiple views of a chequerboard. This process returns a transformation matrix that corrects the distortions in the images captured Szeliski [2022]. For many applications, it is possible to process the images without calibration. It is essential to calibrate images when the geometry is necessary, such as measurements of length and area being made. Entomology work in the field introduces additional factors, including power consumption and storage capacity.

There are three main options for field data collection; digital video recording saved to local storage, network streaming, and live classification. Each different method was examined (see Table 2.2). Local storage and network streaming do not require high computational power and can easily run at camera source resolution and framerate. However, for local storage the data produced will need to be stored on a device which can become a problem for long recording sessions. Network streaming does not have the problem of requiring a large storage device in the field, but it does need a stable network connection and a relatively high network bandwidth to transmit the data to a computer not in the field. Live classification is perhaps the most challenging as while it doesn't require large amounts of storage or a network connection, it does require higher computational power to operate at higher framerates. It also allows a real-time record of pest detections within a preselected area without human intervention. The effect that such a camera trap has on the experiments should be considered though, this is included in the experimental design of the edge computing experiment, Section 2.2.4.

Enhance: Enhancement is a pre-processing step that modifies an image to assist with the segmentation. Modifications to an image made by enhancement include conversion of colour to greyscale (a black and white image), edge detection, filtering, and thresholding (identifying features of an image using their pixel value).

Table 2.2: Resolution, framerate, power consumption, storage requirements, network bandwidth requirements for recording to storage, network streaming and live classification of insects (N = Number of detections, $N > 0$).

	Local Storage	**Network Streaming**	**Live Detection (YOLOv5)**
Framerate(Fps)	30	30	0.2
Potential(V)	5.05	5.03	5.06
Current(A)	0.9	0.79	0.75
Power(W)	4.55	3.97	3.8
Storage Requirements (B)	36,864 * seconds recorded	1,228.8	$\text{floor}(\log_2 N + 1)$
Network Bandwidth (B/s)	0	36,864	0

The filters available include Prewitt, Sobel, Gaussian, Fourier, and Median. The purpose of filters can be to enhance edges and reduce noise. Typically, the Prewitt and Sobel enhance edges, the Gaussian filter is used to blur the image, and the median filter can reduce salt and pepper noise. Fourier filtering allows full control of the spatial frequencies contained in an image. If the camera and scene are not moving, it is possible to use background subtraction to separate the insect from the scene. A codebook algorithm can perform background subtraction in cases where removing the insect from the scene at the start is not possible. Image histogram equalisation modifies the image to equalise the number of pixels for each greyscale value in the image. This enhances the visual contrast but should be used with caution as a pre-processing step in machine learning algorithms as it can reduce the information in the image.

Segment: Segmentation takes an image as input and returns a list of image features. The image features include blobs, lines and contours. A blob is a connected set of pixels in a thresholded image. A Hough transform identifies pixels that belong to lines in the image. The contour describes the bounding pixels of a segmented region. In addition, machine learning methods can be used to segment images semantically, such as the U-Net architecture. The example presented in Section 2.2.4 demonstrates the method.

Feature extraction describes each segmented region of the image. Regions of interest have features which include area, location, eccentricity, and centre of gravity (centroid). These features can be extracted from the image as image moments. The image features can be used to characterise each segmented region. After feature extraction, the image is described by its features rather than its pixels.

Classification reduces the features associated with each segmented region to a single category. Classification provides a semantic label or name (e.g. a species name) associated with the whole image or a region of the image.

Localisation returns the position of the classified feature in the image. The additional step of localisation can be computationally intensive. Typically, localisation returns a bounding box (box that surrounds a classified features) or mask (that can define classified feature exactly) that intersects each instance of the classified object.

To illustrate the machine vision pipeline, a dataset of medically, and forensically important flies Ong and Ahmad [2022] will be used. The dataset's images were captured using a digital single-lens reflex (DSLR) camera (Canon EOS 50D, 15.0 MP APS-C CMOS Sensor) with Tamron 90mm f/2.8 Di Macro. The insects were placed in a lightbox $30 \times 30 \times 30$ cm with $34W$ white light illumination Ong and Ahmad [2022]. The insects were held in place by a pin. Multiple views of the specimens were recorded Ong and Ahmad [2022]. The dataset contains a set of classified images which includes the taxonomic groups *Chrysomya* (731), *Lucilia* (587), *Sarcophaga* (570), Rhiniinae (488), and *Stomorhina* (500). Each RGB image was 224×224 pixels.

The images were **enhanced** by applying classical computer vision methods to obtain a new dataset containing images of the insect specimen with the pin and background removed. In addition, the insect's position within the image was moved to the centre. The following datasets were created: the original image set, a modified image set with pins removed, and a modified re-centred image set with pin and background removed.

To remove the pins in the images, the image's colour was changed to greyscale, then the adaptive thresholding was implemented to separate the insect and the pin from the image background. Hough transformation was applied to detect the line intersecting the pin coordinates for the thresholded image (see Figure 2.8 for a representation of the steps used). A mask the same size as the original image was created where points on the Hough lines were belonging to the pin were set true and all other pixel false. The thresholded image was dilated four times using

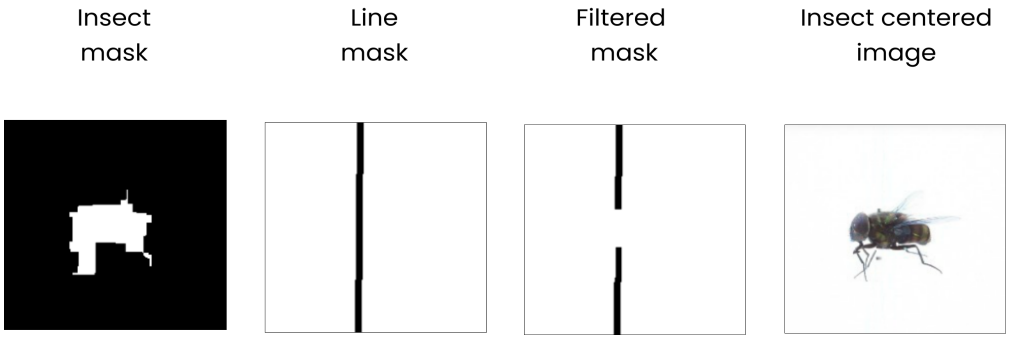

Figure 2.8: The approach used to obtain images of the insects without the pin. The first image contains the mask of the insect, and the second shows the Hough line of the pin. The third image presents the result of the operation combining both masks, and the last image contains the insect with the pin removed.

kernels (matrices of set values applied over an entire image) of size 3×3. This joined regions that were close together. The resulted image was eroded seven times using a kernel of size 3×3. This reduced the size of each feature without reintroducing gaps in the features and also removed the pin.

The region of the eroded image that contains the insect specimen was **segmented** by finding its contours. This operation creates the insect mask, which will later be used to enhance the image. The contours within the bounds expected for an insect were selected to create a mask containing only the insect specimen, M_F (see Figure 2.9). A logical combination of the pin mask, M_L, and the insect mask was used to separate them.

For each pixel location in the altered image, where the pin is true, and the background is false, pixels were transferred to a new reconstructed image using an interpolation algorithm. Working along a line, the pixel values on either side of the pin were identified. Linear interpolation was used to generate pixel values between these two points and add them to the reconstructed image. In all the other cases, the pixel value was transferred from the altered image to the reconstructed image. This removes the pin for regions outside of the insect mask only. The morphological operation of the steps required to remove the pin is described by

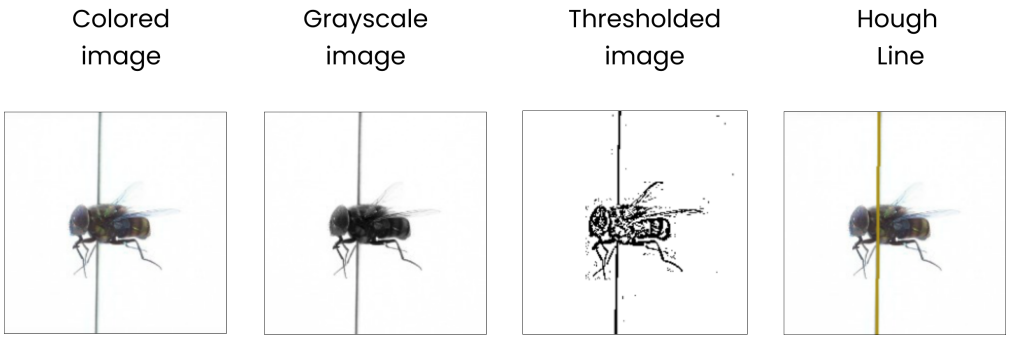

Figure 2.9: Hough lines were used to find and remove the pins in each image of insects. The approach started with the original coloured image. The original image was transformed to a greyscale image. An adaptive thresholding method was then used. Finally, a Hough transformation was applied to the thresholded image to extract the exact position of the pin.

the equation below:

$$I_{NP}(x,y) = \begin{cases} I_P(x,y) & \text{if } M_F \& M_L, \\ I(x,y) & \text{otherwise} \end{cases}$$

where

$$I(x,y) = I_P(x_{min},y)\frac{x - x_{min}}{x_{max} - x_{min}} + I_P(x_{max},y)\frac{x_{max} - x}{x_{max} - x_{min}}$$

is the interpolation operation, I_P is the image of the insect with pin, I_{NP} is the reconstructed image without pin, M_F is the mask of the insect specimen, and M_L the mask of the Hough line. A dataset of insects without pins was created by applying the steps to all images of the original image set. An alternative approach to **segmentation** uses CNN-based methods to classify individual pixels. U-Net is a popular architecture for completing this. The output of this approach is an image with features that can be extracted. This approach is used in Section 2.2.4.

The second dataset of images was created with background pixels removed and the pin removed. The object was translated so that the insect's centre of gravity was positioned at the centre of the image. The translation reduces the classifier's generalisation to detect insects in any image region, therefore is a risk of turning the centre of gravity into a feature the classifier might use. The effect of centring on performance was marginal and is discussed in Section 2.2.3.

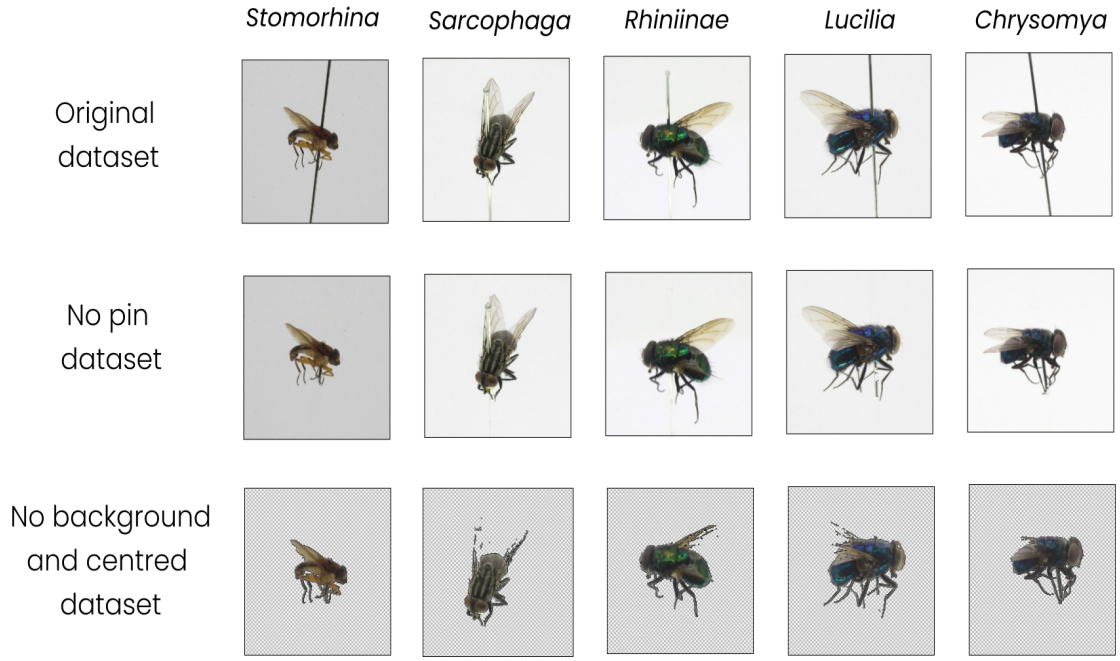

Figure 2.10: Dataset of medically and forensically important flies Ong and Ahmad [2022], and the new datasets created using previously mentioned computer vision techniques. The illustration presents the species taxonomic group in the columns and the dataset type in rows: original, no pin, no background and centred datasets.

After applying the computer vision techniques, three datasets were compiled (see Figure 2.10). To perform the **feature extraction** process, the contours of each image were used to extract its aspect ratio and area of the insect. Coloured images contain Red (R), Green (G) and Blue (B) values for every pixel, the combined value is sometimes referred to as RGB. The average, 2.5% and 97.5% percentile of RGB was obtained from the region which contained the insect in each image. A dataset containing eleven features was assembled. The features included the colour and shape of the insect specimens (see Table 2.3).

To **classify** the flies' taxonomic group, the colour and shape features (see Table 2.3) were used to train a deep neural network algorithm. A convolutional neural network was trained using the fly images (see Figure 2.10). The parameters of a typical CNN were estimated using the TensorFlow package with the Python programming language. In addition, Transfer Learning (TL) concepts were intro-

Table 2.3: A sample of the feature-based dataset obtained from the medically and forensically important flies dataset Ong and Ahmad [2022]. The features obtained was the contour aspect ratio (*Ratio*), contour area (*Area*), average R pixels ($\bar{R}$), average G pixels ($\bar{G}$), average B pixels ($\bar{B}$), 2.5% percentile of R pixels ($R_{2.5\%}$), 2.5% percentile of G pixels ($G_{2.5\%}$), 2.5% percentile of B pixels ($B_{2.5\%}$), 97.5% percentile of R pixels ($R_{2.5\%}$), 97.5% percentile of G pixels ($G_{2.5\%}$), and 97.5% percentile of B pixels ($B_{2.5\%}$). The column Class represents the names given to the taxonomic group of flies (genus or subfamily).

Ratio	Area	$\bar{R}$	$\bar{G}$	$\bar{B}$	$R_{2.5\%}$	$G_{2.5\%}$	$B_{2.5\%}$	$R_{97.5\%}$	$G_{97.5\%}$	$B_{97.5\%}$	Class
0.95	5748	220	219	221	51	56	62	238	238	239	*Lucilia*
1.27	7164	198	194	197	48	36	47	215	215	216	Rhiniinae
0.86	6956	213	211	212	43	37	40	233	233	233	*Chrysomya*
0.41	3500	189	187	188	72	48	63	199	199	199	*Stomorhina*
1.17	7045	208	206	208	52	44	51	226	226	227	*Sarcophaga*

duced using a pre-training CNN architecture called VGG-16, a typical architecture used in TL literature.

2.2.3 Insect Classification using Deep Learning Methods

As mentioned previously, Deep Neural Networks (DNNs) and Convolutional Neural Networks (CNNs) are two deep learning methods. DNNs and CNNs are supervised learning forms, meaning they use labelled data to learn. DNNs are a set of nodes and edges. The nodes are called neurons and each neuron contains a numerical value which is referred to as its activation. Neurons are split into different layers. There are three different types of layers: the input layer, hidden layer and output layer. The input layer takes an input. For example, in an image of 10*10 pixels a DNN would have 100 neurons in the input layer with each neuron representing a greyscale value of each pixel in the image (this operation can be completed with coloured images too). The output layer contains the same number of neurons as possible classes for the image to be classified. A class is usually selected when the neuron in the output layer representing the class has a higher activation than any other neuron in the output layer. The hidden layer consists of one or more layers each comprising of neurons. Layers within the hidden layer don't directly correspond to an input or output but act to convert an input to a correctly classified output.

Every neuron in every layer (excluding the input layer) of a DNN has edges connected to every neuron from the previous layer. These edges are usually referred to as connections which set the activation of the current neuron. The connections contain a weight which is multiplied by each neuron's activation from the previous layer. The weighted value produced by every neuron from the previous layer is then summed and an additional value called 'bias' is added. The value produced could be any real value, but different neuron activations must all be relatively close in value. To fix this the value is inputted into an activation function such as Sigmoid, Rectified Linear Unit (ReLU), Parameterized Rectified Linear Unit (PReLU), Leaky ReLU or hyperbolic tangent (Tanh), which maps values inputted to values within a smaller range. This re-mapped value is set as the current neuron's activation.

DNN is a supervised learning algorithm that estimates their parameters based on labelled data. Let $x_{1,1}, \ldots, x_{i,k}, \ldots, x_{I,K}$ be input variables where I is the number of observations, and K is the number of features and neurons of an input layer from a DNN. Here, these features and labels are presented in Table 2.3. A simple DNN illustration is presented in Figure 2.11 showing an architecture with K input neurons, J hidden neurons and one output neuron. It illustrates the weights $w^h_{k,j}$ and $w^o_{k,j}$ where $j \in \{1, \ldots, J\}$ and $k \in \{1, \ldots, K\}$ for, respectively, hidden and output layers. Also, it shows the biases for both layers θ^h_j and θ^o_1 and the activation function $f(x)$, which can be one of the functions presented in Table 2.4, or others recently proposed Zhang et al. [2021].

Table 2.4: Examples of Activation functions commonly used in Deep learning methods.

Activation Function	**Equation**
Sigmoid	$f(x) = \frac{1}{1+\exp(-x)}$
Tanh	$f(x) = \frac{1-\exp(-2x)}{1+\exp(-2x)}$
ReLU	$f(x) = \max(0, x)$
PReLU	$f(x) = \max(0, x) + \alpha \min(0, x)$
Leaky ReLU	$f(x) = \begin{cases} \max(0, x), & \text{if } x \geq 0 \\ \alpha \max(0, x), & \text{otherwise} \end{cases}$

DNNs utilise linear combinations, which are represented here by

$$net^h_{i,j} = \sum_k^K w^h_{k,j} x_{i,k} + \theta^h_j$$

$$net^o_{i,1} = \sum_k^K w^o_{j,1} f(x_{i,k}) + \theta^o_1$$

for hidden and output layers. The method "learns" by updating the parameters based on, for example, the gradient of a loss function Goodfellow et al. [2016], Mello and Ponti [2018]

$$E^2_i = \left(y_{i,1} - f(net^o_{i,1})\right)^2 = \left(y_{i,1} - f\left(\sum_j^J f(net^h_{i,j}) w^o_{k,j} + \theta^o_k\right)\right)^2.$$

Using a Taylor series expansion, the weights are updated by $w^h_{k,j}(t+1) = w^h_{k,j}(t) - \eta \frac{\partial E^2_i}{\partial w^h_{k,j}}$, $w^o_{k,j}(t+1) = w^o_{k,j}(t) - \eta \frac{\partial E^2_i}{\partial w^o_{k,j}}$, $\theta^h_j(t+1) = \theta^h_j - \eta \frac{\partial E^2_i}{\partial \theta^h_j}$, and $\theta^h_j(t+1) = \theta^o_j - \eta \frac{\partial E^2_i}{\partial \theta^o_j}$, where η is the learning rate. These operations form the basis for the gradient descent algorithm, commonly used to estimate the parameters of DNNs.

The combination of the activation, loss and optimisation functions represents a typical DNN. Nowadays, several combinations of these functions and different numbers of neurons and hidden layers can be used to create a DNN, adding more complexity to the example presented in this chapter Zhang et al. [2021]. A DNN with eleven, ten and five neurons at the input, hidden, and output layers was used to classify the flies' taxonomic group. The leakyRelu activation function was used with $\alpha = 0.3$ in the first two layers. For the output layer, the softmax activation function was selected. The softmax function is a commonly used activation function for converting real-valued vectors into probability distributions. Its output for the presented example is a vector of five elements (each one representing a taxonomic group of fly) between 0 and 1, where the sum of all elements is 1. Each element of this vector will represent the probability the fly belongs to a specific taxonomic group. The cross-entropy loss function and the Adam optimisation algorithm were used to estimate the DNN parameters (see Zhang et al. [2021] for more details).

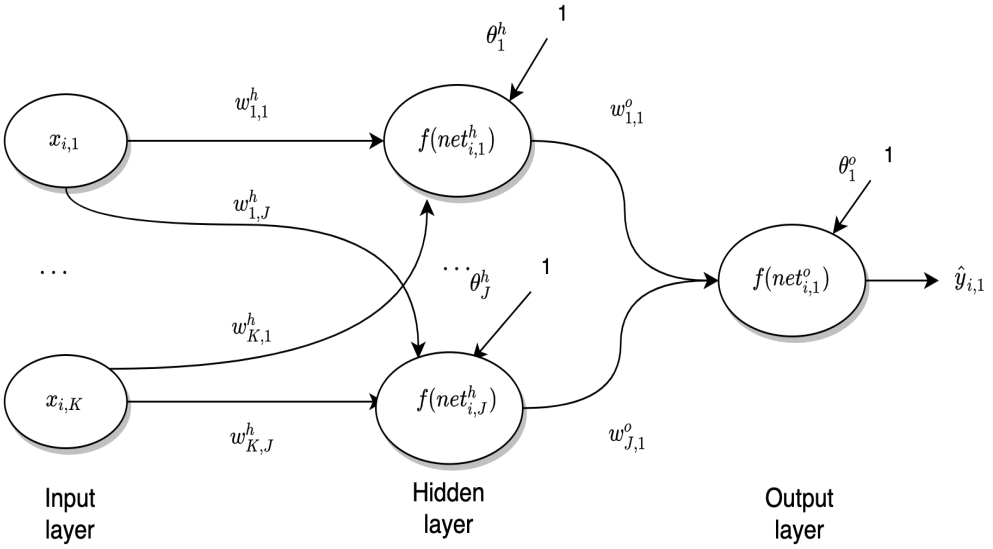

Figure 2.11: Diagram of a simple Artificial Neural Network architecture containing K input neurons, J hidden neurons and 1 output neuron. $w_{k,j}^h$ and $w_{k,j}^o$ are the weights, where $j \in \{1, \ldots, J\}$ and $k \in \{1, \ldots, K\}$ for, respectively, hidden and output layers. θ_j^h and θ_1^o are the biases for hidden and output layers. $net_{i,j}^h$ and $net_{i,1}^o$ represents a linear combinations and f is an activation function. Also, $\hat{y}_{i,1}$ are one-dimensional estimates of the ANN.

A DNN was trained with 70% (2027 observations split into *Chrysomya* (530), *Lucilia* (409), Rhiniinae (341), *Sarcophaga* (398), and *Stomorhina* (349)) of the data for 60 epochs. Its performance was obtained using the unseen complement 848 observations (*Chrysomya* (201), *Lucilia* (177), Rhiniinae (147), *Sarcophaga* (172), and *Stomorhina* (151)). The accuracy statistic was selected to measure the performance of the DNN. In this example, accuracy measures the percentage of correct classifications of flies' taxonomic group by the DNN. Figure 2.12 shows the performance of the DNN when solely using the colour and shape features of the insect flies. The accuracy obtained for the training and validation sets were 75.3% and 59.8%, respectively. In addition, the DNN displayed uneven performance depending on the taxonomic group (26.9% and 1.4% accuracy for predicting the genus *Chrysomya* and Rhiniinae, respectively, while for *Lucilia*, *Sarcophaga* and *Stomorhina* the accuracy was 80.8%, 91.3% and 100.0%, respectively). This indicates the feasibility of using DNNs for classifying these insects based on colour and

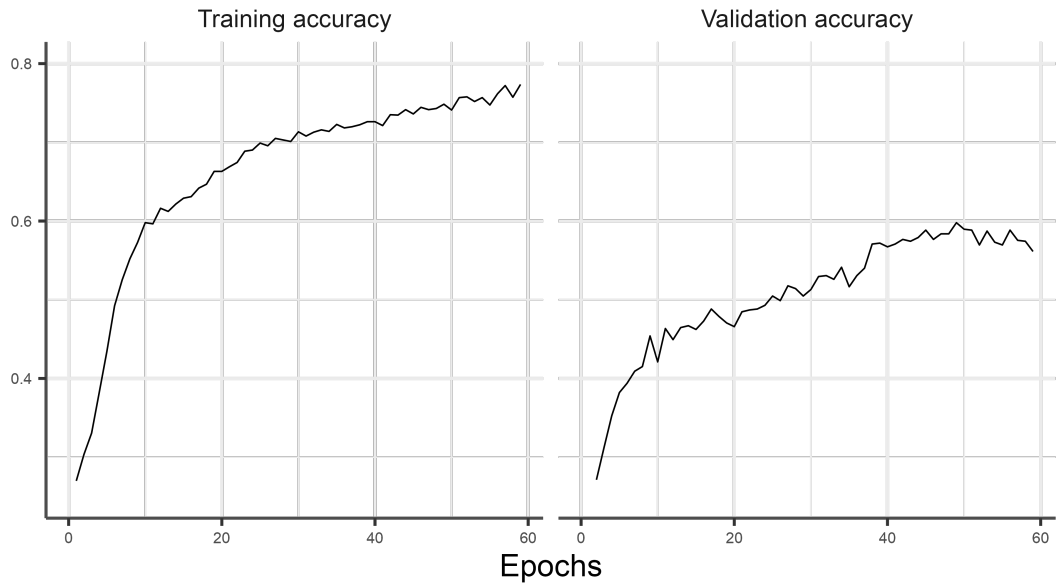

Figure 2.12: Training and validation classification accuracy obtained by the Deep Neural Network used to classify the taxonomic group of medically and forensically important flies based on their colour and shape features.

shape descriptors. However, the method's performance was lower than the validation accuracies usually reported in the literature, considering the classification of other groups of insects. To improve the classification performance a convolutional neural network was implemented to classify the flies. This method creates features with more information than the features used for the DNN example.

Convolutional neural networks are a particular type of deep neural network for processing data with a known, grid-like topology, such as time series and image data Goodfellow et al. [2016]. CNN is a supervised learning method that is given a group of images including their labels; in this case, images of flies accompanied by their taxonomic group. We estimate the parameters of the CNN using a subset of the overall data, and calculate the performance of the method using the unseen images.

Once an image is presented to a CNN, the initial layers extract features that allow

insect classification. The feature extraction is performed by adding convolutional and subsampling layers to a DNN architecture Goodfellow et al. [2016], Wang et al. [2019], Zhang et al. [2021]. A typical CNN contains convolution, subsampling, and fully connected layers. The fully connected layers are simply layers of a DNN, indicating that the "learning" process also uses the presented activation, loss and optimisation functions used to estimate the parameters of a DNN. The CNN architecture contains more parameters to estimate than a DNN, mainly because of the convolution layers. This layer uses an operation called convolution that is defined by the equation below:

$$C_{r,q}^{d} = f\left(\sum_{m=1}^{M}\sum_{n=1}^{N} w_{m,n}^{c} x_{r+m-1,q+n-1} + \theta^{c}\right),$$

where $x_{r,q}$ are pixels of a greyscale image $\mathbf{X}$ of size $R \times Q$, $w_{m,n}^{c}$ and θ^{c} are hyperparameters of a filter $\mathbf{W}$ with size $M \times N$, f can be any activation function, and (r, q) are the indices of the output, also known as a *feature map*. This operation can be repeated d times per convolution layer. Figure 2.13 shows an example of the convolution operation with one filter of size 3×3. The optimisation algorithms used in this architecture will find the values of each $w_{m,n}^{c}$ and θ^{c} for all the filters used in each convolution operation added in the CNN architecture in such a way that the classification error is reduced. Multiple optimisation algorithms were created to achieve this task in the most efficient way Zhang et al. [2021].

Typically since small filters are used in convolution operations, a few pixels may be lost. A solution for this problem is the padding approach, which consists of adding extra pixels around the boundary of the input image, increasing the effective size of the image Zhang et al. [2021]. These extra values are typically zero. Figure 2.14 shows the result of a convolution operation with padding and explains how the size of an image can be maintained after a convolution operation.

The subsampling layer contains operations that significantly reduce the size of the image. These operations are essential to create the feature vector that carries the information of the presented image. There are two commonly used types of subsampling operations, max pooling and average pooling. A max pooling operation with kernel size 3 is defined below:

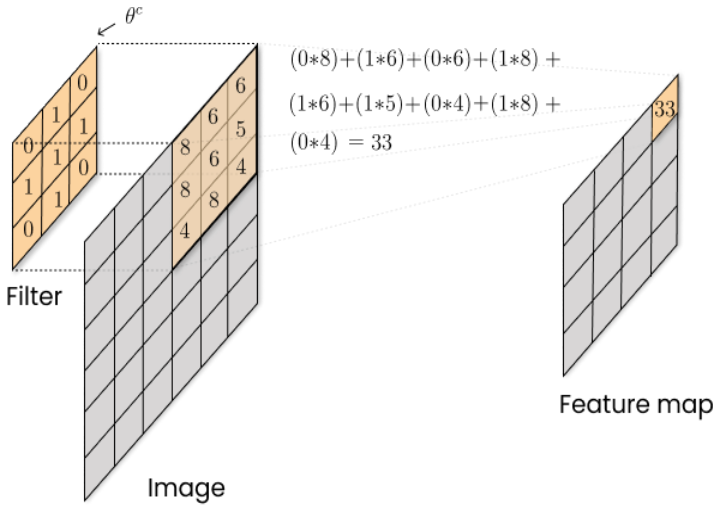

Figure 2.13: Illustration of a convolution operation using a filter of size 3×3 on a greyscale image with indices (r, q). The parameter $\theta^c = 0$ for this example, and the feature map presents the results of the first iteration of convolution using a ReLu activation function, $\max(0, x)$.

$$P^d_{r,q} = \text{MaxPool}(C^d_{r,q}) = \max \begin{pmatrix} C^d_{r,q} & C^d_{r+1,q} & C^d_{r+2,q} \\ C^d_{r+1,q} & C^d_{r+1,q+1} & C^d_{r+2,q} \\ C^d_{r+2,q} & C^d_{r+2,q+1} & C^d_{r+2,q+2}, \end{pmatrix}$$

where (r, q) are the indices of the dth feature map of a previous convolution layer. Figure 2.15 shows an illustration of a max pooling operation with a kernel size of 3 on a greyscale image of size 6×6 pixels.

A CNN with six convolutional, five max pooling and three fully connected layers was created to classify the flies' taxonomic group. The leakyRely activation function was used with $\alpha = 0.3$ for all convolutional layers and the first two fully connected layers. For the last fully connected layer, the softmax activation function was used. In addition, the cross-entropy loss and the Adam optimisation function were used to estimate the parameters of the CNN. Figure 2.16 presents a diagram of the proposed CNN architecture. The CNN was trained using 70% (2027 images split into *Chrysomya* (530), *Lucilia* (409), Rhiniinae (341), *Sarcophaga* (398), and *Stomorhina* (349)) of the dataset for 30 epochs. The method's performance was obtained using 30% of the dataset (848 images containing *Chrysomya* (201), *Lu-*

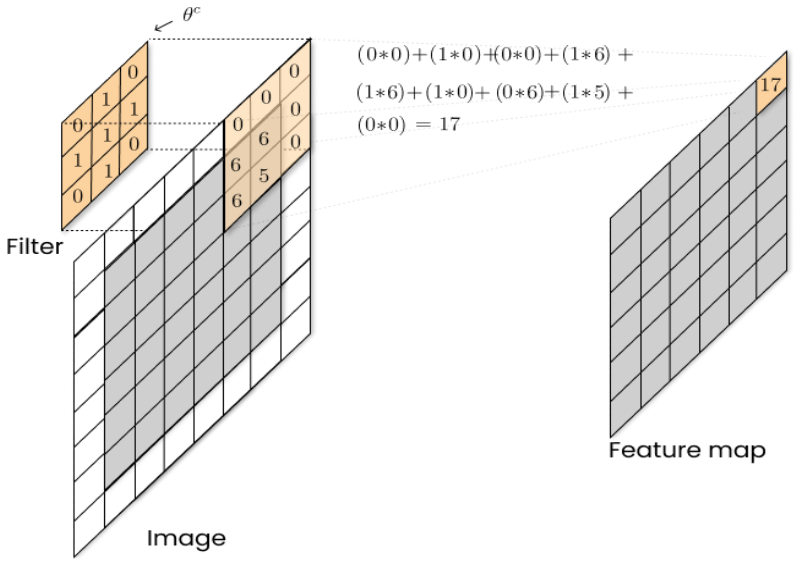

Figure 2.14: Illustration of a convolution operation with padding using a filter of size 3×3 on a greyscale image with indices (r, q). The parameter $\theta^c = 0$ for this example, and the feature map presents the results of the first iteration of convolution using a ReLu activation function, $\max(0, x)$.

cilia (177), Rhiniinae (147), *Sarcophaga* (172), and *Stomorhina* (151)) based on the accuracy statistic. The performance of the CNN was obtained for all image sets (see Figure 2.10).

Figure 2.17 shows the training and validation accuracy obtained by the CNN architecture for each image set. The best performance obtained using the original flies dataset was 100.0% and 75.2% accuracy for the training and validation set, respectively. The CNN presented different accuracies per group, namely *Chrysomya* (16.4%), *Lucilia* (91.0%), Rhiniinae (85.0%), *Sarcophaga* (99.4%), and *Stomorhina* (98.0%). The best performance obtained using the dataset with images without the pin was 99.5% and 76.3% accuracy for the training and validation sets, respectively. The taxonomic group accuracies were: *Chrysomya* (25.4%), *Lucilia* (88.7%), Rhiniinae (86.3%), *Sarcophaga* (93.6%), and *Stomorhina* (100.0%). In addition, the best performance obtained with the image set without pin, background and centred was 98.5% and 75.3% for training and validation, respectively. The accuracies were: *Chrysomya* (16.4%), *Lucilia* (97.7%), Rhiniinae (78.2%),

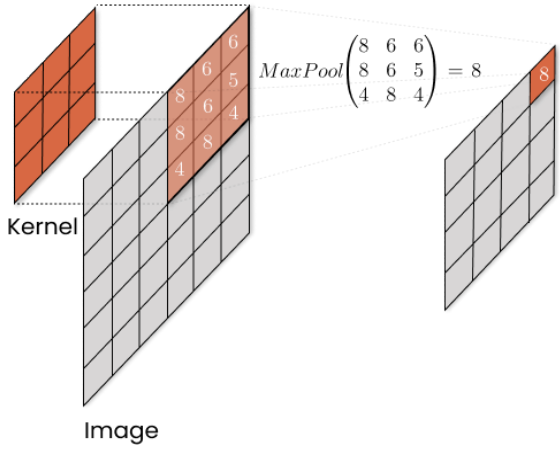

Figure 2.15: Illustration of the max pooling subsampling operation using a kernel of size $3x3$ on a feature map of size 6×6. The output of this operation is a matrix of size 4×4.

Sarcophaga (100.0%), and *Stomorhina* (96.7%). This indicates a significant increase in performance compared to the DNN architecture presented earlier. The convolutional and max pooling layers provided extracted features from the insect's image that improved classification performance.

The first fully connected layer can be seen as the features presented in Table 2.3. However, these features are not restricted to statistics used to produce Table 2.3. These values are obtained from the convolutions and subsampling operations used in the CNN, which are difficult to interpret. On the other hand, CNNs can provide higher classification performance indicating their potential for insect classification tasks. The example presented illustrates the potential of CNNs for classifying insects based on their images.

Another approach to training a CNN is based on Transfer Learning (TL). TL uses a trained CNN architecture to extract the features from images with new or already seen classes. TL assumes that the parameters estimated by a CNN using millions of images with various classes provide a generic model that can be used for other problems. A commonly used dataset to train CNNs is 'ImageNet', which contains $14,197,122$ images of many different classes. Several authors used the ImageNet dataset to propose CNN architectures, such as VGG16, AlexNet, LeNet,

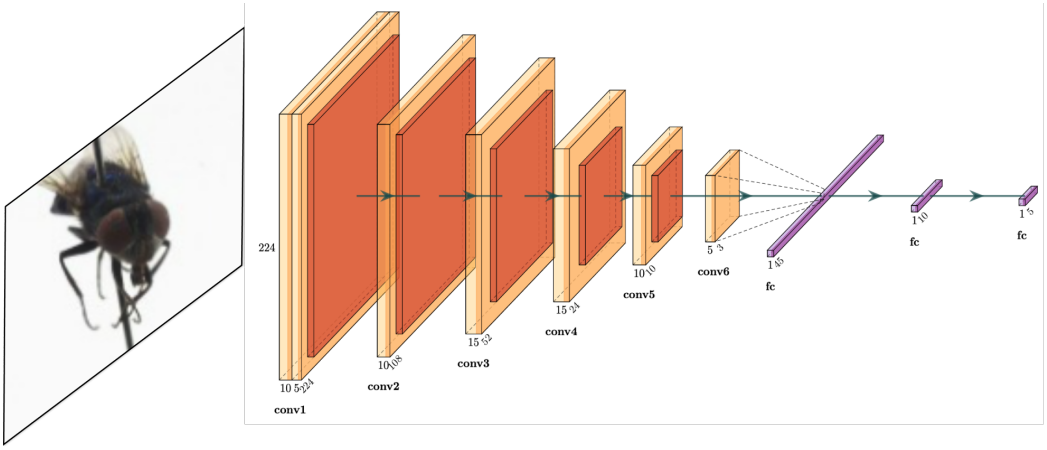

Figure 2.16: Convolutional neural network architecture used to classify insect's taxonomic group. Orange matrices are convolutional layers (conv) whose depth is the number of filters. Red matrices represent the max pooling operation applied on the feature maps. Magenta vectors represent the fully connected layers (fc). This diagram was created with the package PlotNeuralNet Iqbal [2018].

EfficientNet and others Wani et al. [2020], Zhang et al. [2022]. A higher increase in CNN performance by applying TL to their problems was reported multiple times Goodfellow et al. [2016], Wang et al. [2019]. In a nutshell, TL can improve the performance of classification tasks by providing a general feature extractor based on typical CNN operations Goodfellow et al. [2016], Chollet [2018], Wani et al. [2020], Zhang et al. [2022].

A VGG16 architecture was implemented to classify the flies' taxonomic group. An illustration of the VGG16 architecture is presented in Figure 2.18. TL was implemented using a pre-trained VGG16 architecture based on the ImageNet dataset. Three fully connected layers were added to the architecture, containing $25,088$, 10 and 5 neurons, respectively. The first layer is obtained by flattening the output of the last max pooling operation produced by the VGG16 architecture. As this output is a tensor of size $(7, 7, 512)$, the first fully connected layer will have $7 * 7 * 512 = 25,088$ neurons. Training and validation sets with 70% and 30% of the dataset, respectively, were used. The leakyRelu activation function was used with $\alpha = 0.3$ for the first two layers, and a softmax activation function was used in the last layer. Using a cross-entropy loss and Adam optimisation function, the

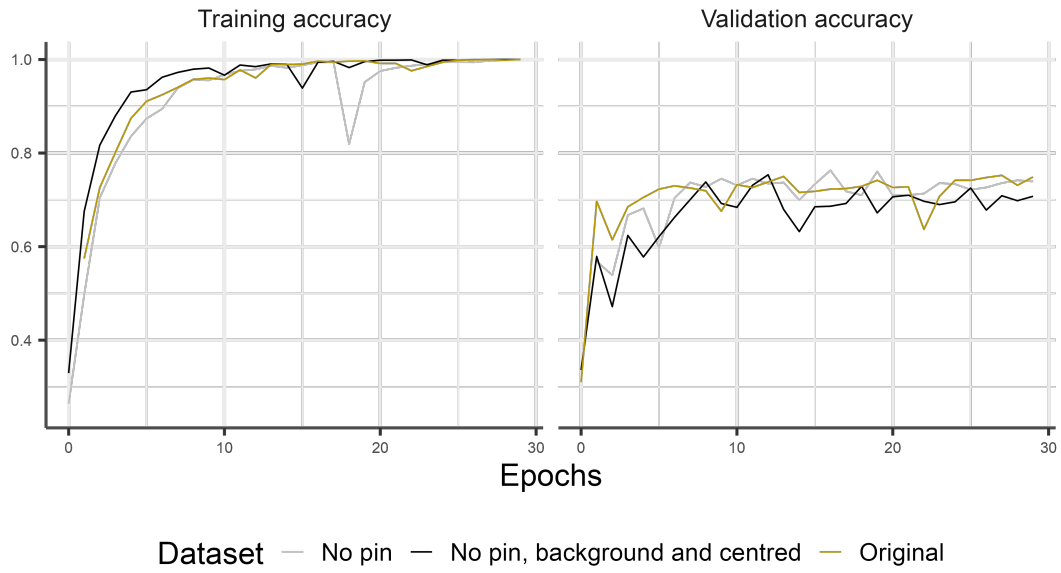

Figure 2.17: Train and test classification accuracy obtained by the convolutional neural network to classify taxonomic group of medically and forensically important flies based on original, no pin and no pin, background and centred datasets.

parameters of the fully connected layers were optimised with the training data using 30 epochs. The performance of the architecture was assessed using accuracy measures for the validation data.

Figure 2.19 presents the training and validation accuracy based on the pre-trained VGG16 architecture. A training and validation accuracy of 92.6% and 76.5% were obtained for the original flies dataset, with group accuracies: *Chrysomya* (40.3%), *Lucilia* (89.3%), Rhiniinae (65.3%), *Sarcophaga* (99.4%), and *Stomorhina* (94.7%). A training and validation accuracy of 100.0% and 74.1% were obtained based on the flies dataset without pins, with group accuracies: *Chrysomya* (35.3%), *Lucilia* (56.5%), Rhiniinae (57.5%), *Sarcophaga* (98.2%), and *Stomorhina* (100.0%). The performance for training and validation was 100.0% and 67.8% for the flies dataset without background, pin and specimens centred, with group accuracies: *Chrysomya* (35.8%), *Lucilia* (72.3%), Rhiniinae (73.5%), *Sarcophaga* (98.8%), and *Stomorhina* (100.0%). The pre-trained VGG16 architecture converges to a higher

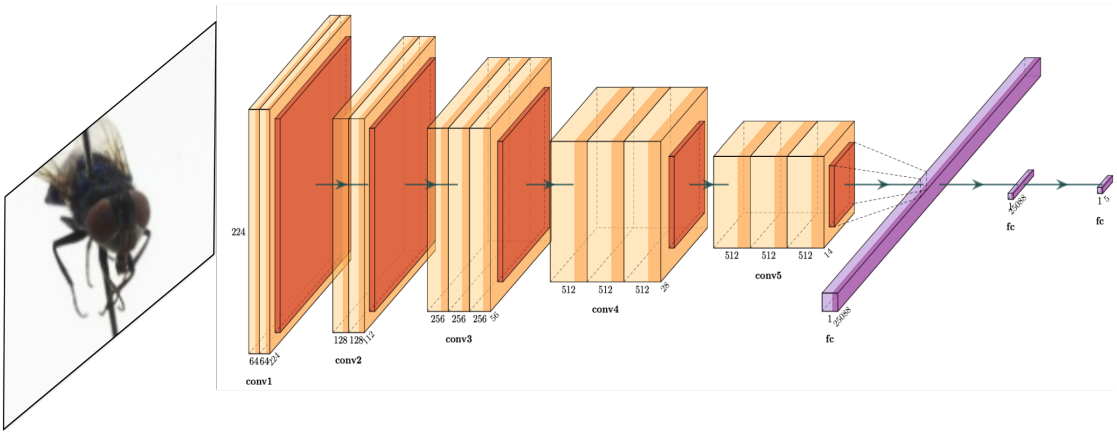

Figure 2.18: VGG16 architecture. Orange matrices are convolutional layers (conv) whose depth is the number of filters. Red matrices represent the max pooling operation applied on the feature maps. Magenta vectors represent the fully connected layers (fc). This diagram was created with the package PlotNeuralNet Iqbal [2018].

training accuracy faster than the previous CNN architecture, and after 25 epochs, both architectures' training performances converged to values close to 100.0%.

The validation performance of the VGG16 architecture and the proposed CNN architecture using the original and no-pin image sets are similar. There is a quicker convergence of the validation accuracy value in the VGG16 architecture compared to the CNN trained from scratch. On the other hand, for the dataset of images without background, pin and insect specimens centred, the VGG16 architecture obtains a lower validation accuracy than the proposed CNN architecture. This indicates that the parameters optimised with the ImageNet dataset did not provide a better generalisation for classifying medically and forensically important flies' taxonomic group compared to optimising all CNN layers based on the available images.

The datasets presented here are high-dimensional. For instance, the colour-based dataset of medically and forensically important flies (see Table 2.3) contains 11 dimensions, and the image-based dataset using the proposed CNN has 45 dimensions. Using transfer learning based on VGG16 architecture, the image-based dataset has $25,088$ dimensions. Higher-dimensional datasets are commonly found

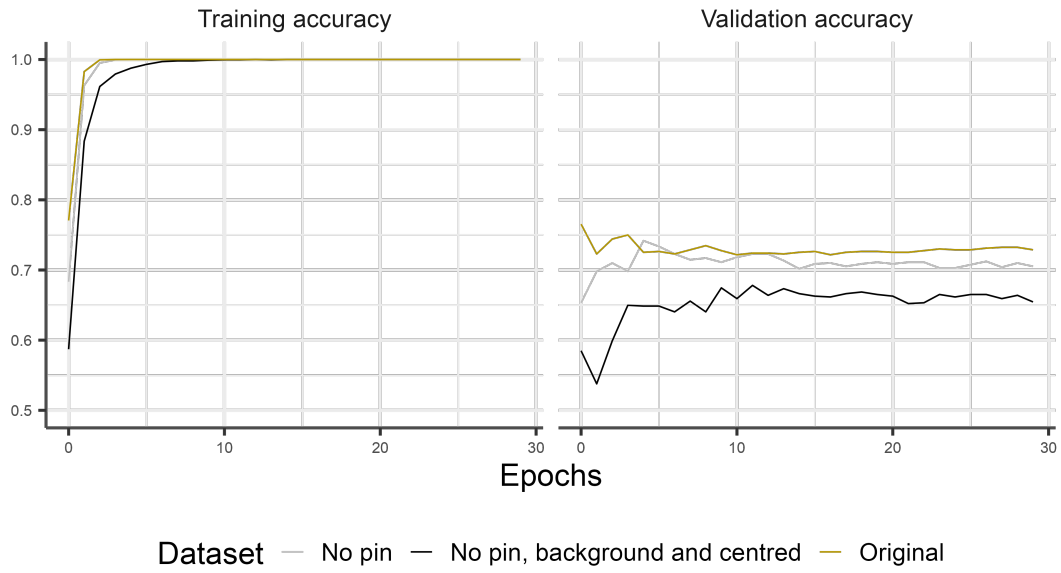

Figure 2.19: Training and test classification accuracy obtained by the pre-trained VGG16 architecture to classify taxonomic group of medically and forensically important flies based on original, no pin and no pin, background and centred datasets.

in machine vision problems in general, so Dimensional Reduction (DR) techniques are essential for understanding datasets' geometric and neighbourhood structure, and further interpreting the CNN performance Wang et al. [2021]. Several unsupervised learning methods can be used for DR tasks, such as t-SNE, UMAP, TriMap, and PaCMAP Wang et al. [2021].

The available DR methods can be clustered into local and global methods. The main difference between them lies in whether they aim to preserve global or local structure in the high-dimensional data Wang et al. [2021]. To clarify the difference between local and global DR methods, the analogy based on a dataset of planets and moons from a planetary system introduced by Wang et al. [2021] is presented. The aim of local methods is to preserve relative distances between moons and planets without considering relative distances between planets. On the other hand, the main goal of global methods is primarily to preserve the relative distances of planets without the consideration of the positions of moons around planets.

To visualise the geometric and neighbourhood structures of the medically and forensically important flies datasets used for classification, the Pairwise Controlled Manifold Approximation Projection (PaCMAP) method Wang et al. [2021] was used, considering its capability to preserve both local and global structures. Also, the algorithm has fast convergence thanks to its simple loss function Wang et al. [2021]. Thus, PaCMAP was applied to the dataset containing colour-based features from the flies specimens. Also, it was applied to the dataset of features obtained from the pre-trained VGG16 architecture using the flies' images. Finally, PaCMAP was applied to the dataset of features obtained from the proposed CNN architecture using the flies' dataset. All the observations of each dataset were used in the analysis.

Figure 2.20 presents the lower dimension dataset using the PaCMAP method applied to the medically and forensically important flies dataset based on colour-based features extracted from flies. The visualisation presents three central clusters, indicating that using only colour-based features is insufficient to produce distinct clusters for each flies' class. An aggregation of observations related to the genus *Stomorhina* is presented in Figure 2.20, indicating a presence of a local structure on the high-dimensional dataset. As stated before, the percentage of correct classification by the DNN for *Chrysomya*, *Lucilia*, Rhiniinae, *Sarcophaga*, and *Stomorhina* were respectively $26.9\%, 80.8\%, 1.4\%, 91.3\%$ and 100.0%. The DNN obtained higher correct classifications solely in three fly genus, emphasising that colour features are not enough for classifying all these genera.

Applying the PaCMAP method on the features extracted by the proposed CNN trained with the original, no pin and no pin, background and centred datasets is presented in Figure 2.21, showing an improvement in the number of clusters compared to the colour-based features. For all datasets, it is possible to distinguish between the five classes considering that there is less overlapping among the observations. An expressive increase of correct classification for the subfamily Rhiniinae was observed using the features from the proposed CNN. The visualisation suggests a misclassification between the genera *Chrysomya* and *Lucilia*. The fact that the CNN misclassified 50.2%, 36.8%, and 53.7% of the specimens' images of the genus *Chrysomya* as *Lucilia* for respectively the original, no pin, and no

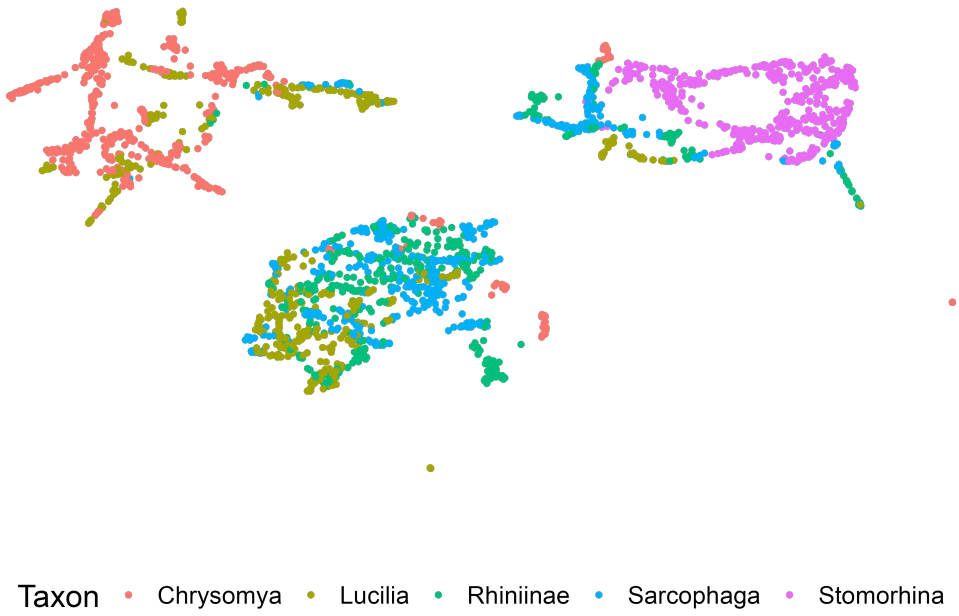

Figure 2.20: Application of PaCMAP to the medically and forensically important flies dataset based on colour features of the flies' specimens using computer vision techniques. The colour represents each taxonomic group of flies presented in the dataset.

pin, background and centred datasets supports PaCMAP visualisation.

Figure 2.22 shows the PaCMAP visualisation of the extracted features by the pre-trained VGG16 CNN for all the datasets presented here. We observe three clusters and other small groups for the original dataset. However, the observations related to the genera *Sarcophaga*, *Lucilia*, and subfamily Rhiniinae overlap in the visualisation. The 89.5%, 99.4%, and 94.7% of correct classification obtained for the genus *Lucilia*, *Sarcophaga*, and *Stomorhina* indicates the structure of the features provided by the VGG16 architecture caused a reduction of classification performance for the specimens of the subfamily Rhiniinae. For the no-pin dataset, observing three main clusters and smaller clusters of specimens of the genus *Lucilia* is possible. It indicates that the reduction of performance of the CNN can be caused by the complexity of the features extracted by the VGG16, which provides feature structures overlapping the observations related to the genera *Lucilia*, *Sar-*

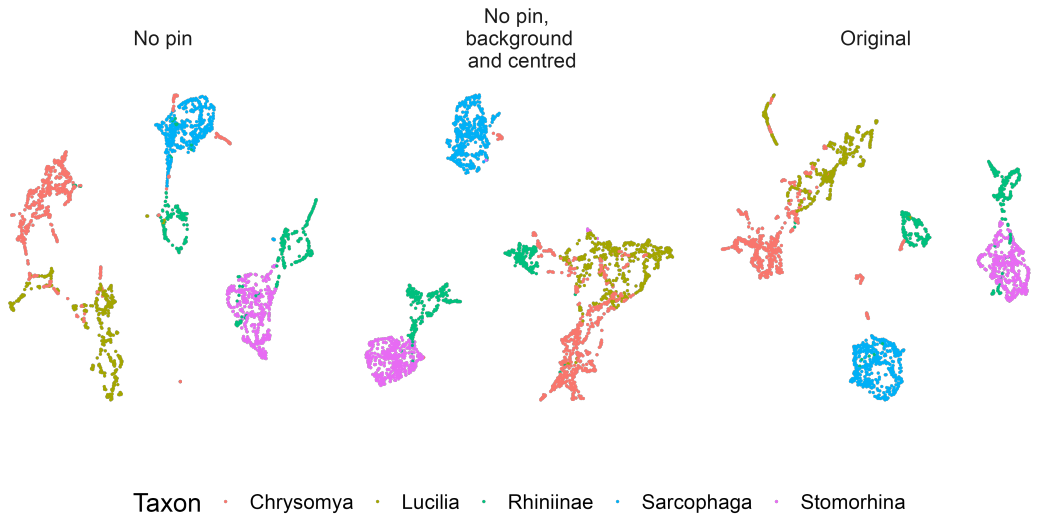

Figure 2.21: Application of PaCMAP to the medically and forensically important flies datasets based on flattened original images of flies, images of flies with no pin, and no pin, background, and centred. The colour represents each taxonomic group of flies presented in the dataset.

cophaga, and the subfamily Rhiniinae. Finally, for the non-pin, background and centred dataset, it is possible to observe three central clusters. However, they are connected by some observations and overlapping is frequently found based on the PaCMAP visualisation. It indicates that the features obtained by the pre-trained VGG16 architecture are insufficient to provide a higher classification accuracy for this dataset when compared to the proposed CNN trained with this dataset.

Overall, using dimensional reduction methods that allow us to observe a high-dimensional dataset's geometric and neighbourhood structures allows us better understand the classification performance of machine learning methods. Therefore, using these methods in the context of machine vision, where high dimensional datasets are commonly encountered, helps researchers to obtain more insights about their models. In addition, combining entomologists' expertise when pre-processing the insects' images allows for better interpretation of the problem.

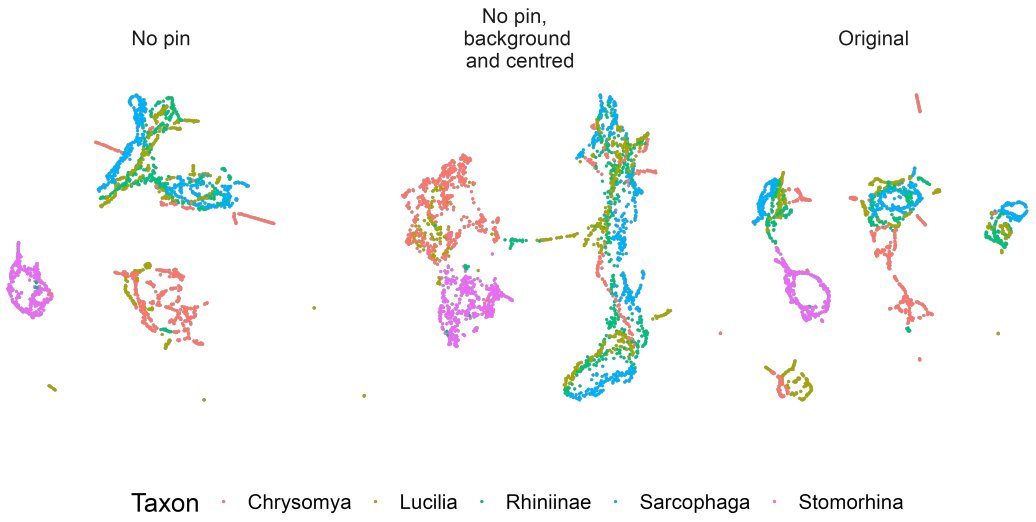

Figure 2.22: Application of PaCMAP to the medically and forensically important flies datasets based on the VGG16 feature extracted from original flies' images of flies with no pin, and no pin, background, and centred. The colour represents each taxonomic group of flies presented in the dataset.

2.2.4 Insect localisation with deep learning methods

A common approach used in CNN-based research is visualising regions of an image that most contribute to the output of a classifier. Multiple authors have proposed methods to find these regions, including Gradient-weighted Class Activation Mapping (Grad-CAM), Grad-CAM+, Grad-CAM++, and Salience maps Selvaraju et al. [2017], Alqaraawi et al. [2020], Lerma and Lucas [2022]. These methods are widely used for the localisation of objects in an image. The Grad-CAM method is often reported in CNN-based papers Palma et al. [2020], Lerma and Lucas [2022], Setiawan and Rulaningtyas [2023]. The Grad-CAM method uses the gradient of the last convolution layer of the VGG16 architecture to obtain the regions of the image that contributed to the classification. This method was implemented with the medically and forensically important flies dataset using the Pre-trained VGG16 architecture to illustrate its application on insect localisation.

Figure 2.23 shows the heatmap produced by the method. The VGG16 architecture classified the image as a 'mosquito' based on the ImageNet dataset, and the

heatmap presents the regions of the image that contributed to this classification. Using the heatmap, it is possible to see that the pin is an important feature of this classification, considering the original dataset. When the pin is removed, only the pixels related to the insect specimens contribute to the classification.

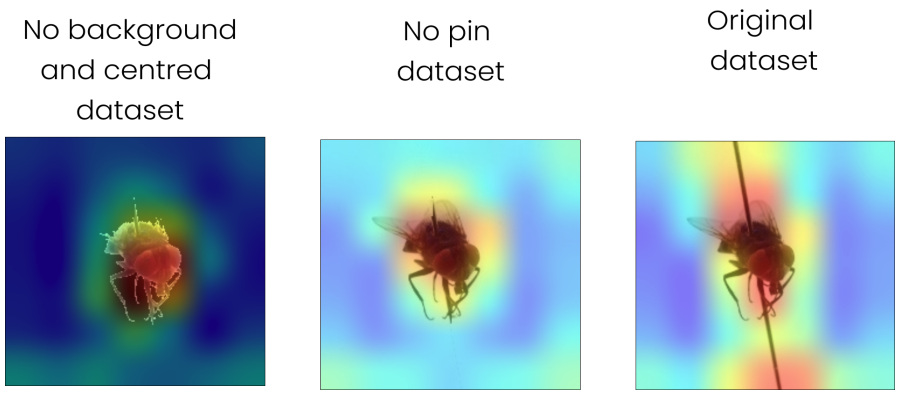

Figure 2.23: Visualisation of the feature extraction based on the Gradient-weighted Class Activation Mapping (Grad-CAM). The heatmap represents the weighted channels at the feature map by the gradient provided by the class 'mosquito' of the ImageNet dataset.

The Grad-CAM can be used to analyse the presence of bias in a CNN. Grad-CAM heatmaps that present regions not belonging to the target classification indicate that the CNN is using not relevant features for the classification. The example presented illustrates this point. The pin is not a feature of the flies' taxonomic group, however, the pre-trained VGG16 is using it in the original dataset to assist with classification.

An alternative method for localising insects in an image is the Unet architecture. The Unet architecture consists of a contracting path based on the typical operations of a CNN. Also, an expansive path based on upsampling the number of feature channels and concatenation with the correspondingly cropped feature map from the contracting path is part of the Unet (see Figure 2.24). The U-shape is presented on the architecture's name because the original image is contracted to a lower size and expanded to a higher image size Ronneberger et al. [2015a]. The Unet requires labels containing the regions belonging to the target class, and the labelling process is usually done manually.

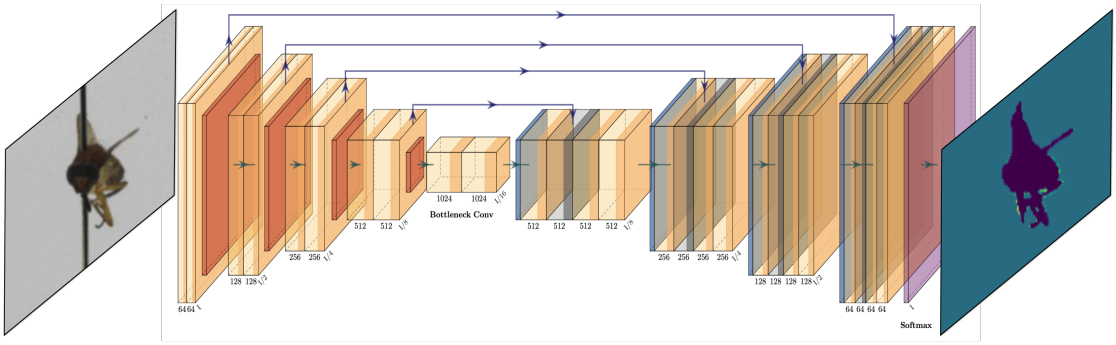

Figure 2.24: Unet architecture used to segment insect's taxonomic group in the image. Orange matrices are convolutional layers (conv) whose depth is the number of filters. Red matrices represent the max pooling operation applied on the feature maps. Blue matrices represent the transpose convolution layer. The arrows represents the concatenated filter maps obtained from the contracting path to the expanding path. The purple matrix represents the tensor output of size $(224 \times 224 \times 6)$, where each of the 6 channels represents the background, and the five flies' taxonomic groups. This diagram was created with the package PlotNeuralNet Iqbal [2018].

Computer vision methods were used to obtain the segmented labels for the segmentation task. The original dataset of medically and forensically important flies was labeled based on contour detection. The labels are images of size 224×224 with only one channel, where the pixels represent the class found on the original image. The pixels related to the background were equal to 2, and the pixels related to, respectively, the groups *Stomorhina*, *Sarcophaga*, Rhiniinae, *Lucilia*, and *Chrysomya* are equal to 3, 4, 5, 6, and 7. Unet was implemented with five blocks on the contracting path, where each block contained a convolution, max pooling, dropout and batch normalisation operations. A 10% dropout operation was added to the architecture, meaning that each neuron had a 10% probability of being deactivated. Batch normalisation was used to normalise a batch of 32 images. Four additional blocks were used in the expansive path, where the upsampling version of convolution, transpose convolution, and max pooling were added.

The last layer of the presented Unet contains a convolution layer with a softmax activation function. The softmax allows the segmentation of multiple classes. The presented example has six classes: the background and five taxonomic groups of flies. Therefore, the output of the Unet will be an array of size (224×224) with six channels, each related to a specific class. Finally, we used a sparse categorical cross-entropy function loss and the Adam algorithm to perform the optimisation of the parameters of the Unet. 70% of the original dataset was used for training with 30 epochs to estimate the parameters of the Unet, and 30% of the original dataset was used for validation. Again, we used accuracy as the performance metric.

Figure 2.25 shows the training and validation accuracy of the Unet architecture. The best performance obtained was 94.4% and 92.8% accuracy for training and validation, respectively. Figure 2.24 illustrates the Unet architecture and shows the segmented output of an image containing a specimen of the genus *Stomorhina* as an example. The selected image of the fly is part of the validation set, and the yellow colour presented in the segmentation indicates pixel-wise classification errors provided by the Unet. Overall, the method produces a promising result by segmenting the region where the specimen is found and correctly classifying the regions of the image in the majority of the image's pixels.

Another approach for localising insects in images is finding a bounding box containing the target insect. One of the most used CNN-based methods for achieving this goal is the 'You Only Look Once' (YOLO) architecture. YOLO is an architecture that classifies and localises objects within an image, built on a CNN that simultaneously predicts multiple bounding boxes and class probabilities Redmon et al. [2016]. A dataset of labelled images is required to train a model using YOLO. This dataset must comprise of both a set of images that contain the objects of interest and a set of labels. Every object must appear at least once in the set of images, but an increase in the number of appearances and different orientations of an object increases the accuracy of the YOLO model. The labels in the dataset specify what objects appear and where they appear in each image. The images and their accompanying labels can then be used to train a YOLO model. Once trained, the YOLO model will be able to classify and localise new images that contain the objects of interest. YOLO is an architecture with many different versions such as

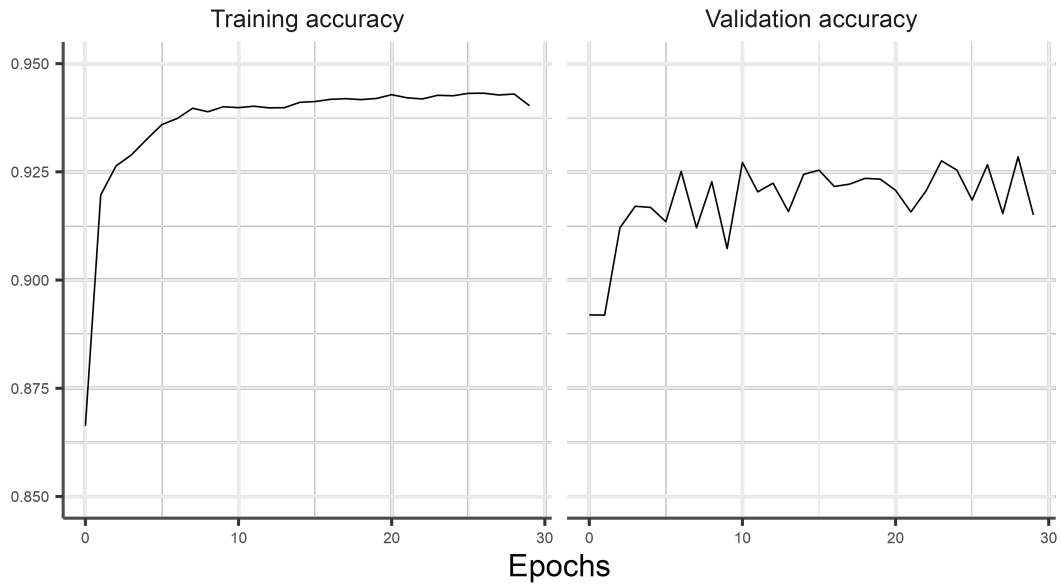

Figure 2.25: Training and test classification accuracy obtained by the Unet architecture to classify taxonomic groups of medically and forensically important flies based on the original dataset with segmented labels.

YOLOv5 (https://github.com/ultralytics/yolov5) that are readily available and provide a great starting point for anyone interested in performing automated detection.

2.2.5 Platforms of Computing

To expand the possible methodologies that can be used in the entomological context. We now present another tool that enables insect localisation in more complex backgrounds. This is relevant because live monitoring poses challenges due to background complexity, and more specialised methods are required to improve performance [Zhang et al., 2023b, Ștefan et al., 2025].

This example presents a challenge to the studied Caughley's management action. This challenge is not restricted to insect monitoring. However, here we present a proof-of-concept example to illustrate the potential of YOLOv5 for insect lo-

calisation and, with further implementation, for counting and classifying these specimens in the field [Venverloo and Duarte, 2024]

As previously mentioned, to perform automated live experiments in the field multiple factors need to be considered. These include the detection algorithm, computing device, framerate, power consumption, and operating temperature of the device. There are many methods to automate the detection of insects within an image. For this experiment YOLOv5 was selected, as it is effective at classifying objects and allowed for the localisation of objects to bounding boxes within an image. The ability to classify and localise objects allows for multiple objects to be counted within a single frame of detection.

Data to train a YOLOv5 architecture was gathered. The common wasp *Vespula vulgaris* was chosen as a sample insect for detection due to its high abundance in Ireland throughout the warmer months. Images of *Vespula vulgaris* visiting a feeding station were recorded in both August and September of 2022 to train the model. The images were manually labelled using the online tool 'Make Sense' (`https://www.makesense.ai/`). This process can also be completed offline with a text editor.

After the data was labelled, it was spilt into training and validation sets comprising of 35 and 17 labelled images, respectively. The data sets were used to train a model on the YOLOv5n6 architecture (one of the many variations of YOLOv5), which produces both reasonably accurate and very fast models. The model was generated using the Ultralytics Hub web application (`https://ultralytics.com/`). The resulting model was able to classify and locate *Vespula vulgaris* within images (see Figure 2.26).

The detection model was tested on multiple devices. As the model was generated on the Ultralytics Hub web application it was deployable to a smartphone. To retain consistency while testing the other devices, the model from the web application was exported to the various required file formats for the other devices. Each device (excluding the mobile phone and the Luxonis Oak-1-PCBA) ran the model

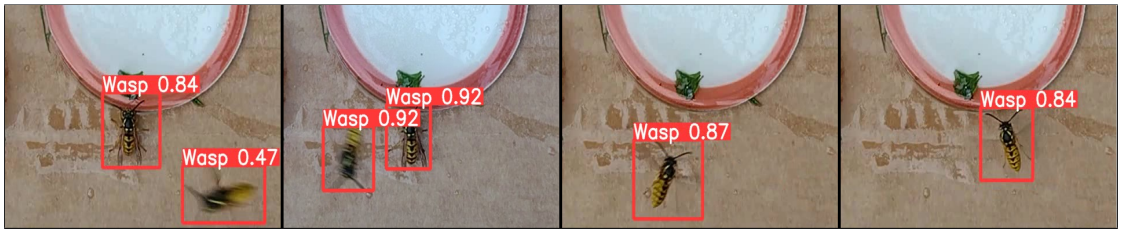

Figure 2.26: *Vespula vulgaris* classified and localised by YOLOv5 in four images.

using the same detection script from a locally downloaded version of YOLOv5 [2]. The detection script was altered to display an average framerate in the top left corner of the processed frames during live detection [3]. This may have lowered the framerate slightly for the devices. Framerate for the mobile phone was recorded using an external application [4]. The models were run on a Laptop with an Intel Core i7-10870H CPU @ 2.20GHz CPU and a NVIDIA GeForce RTX 3060 Laptop GPU (see Figure 2.27 A), a Samsung Galaxy S20 FE 5G Mobile Phone (see Figure 2.27 D), Raspberry Pi (see Figure 2.27 F), Intel Neural Compute Stick 2 (see Figure 2.27 H), Coral USB Accelerator (see Figure 2.27 J), and a Luxonis Oak-1-PCBA (see Figure 2.27 M).

The Raspberry Pi 4 8GB RAM Model B CPU (Broadcom BCM2711, Quad core Cortex-A72 (ARM v8) 64-bit SoC @ 1.5GHz) (see Figure 2.27 F) is a small single-board computer with video output network and USB support, capable of performing the same tasks as most computers. For the experiment the Raspberry Pi used the 32-bit Raspbian (Bullseye) OS to ensure compatibility with a Raspberry Pi High Quality Camera. As of writing it was difficult to get the 64-bit Raspbian OS working with the high-quality camera [5]. This later proved to be an issue as YOLOv5 is dependent on a machine learning framework called PyTorch [6]. Py-

[2] https://github.com/ultralytics/yolov5/blob/master/detect.py. Accessed: 2025-01-08

[3] https://github.com/ultralytics/yolov5/discussions/6713. Accessed: 2025-01-08

[4] https://play.google.com/store/apps/details?id=com.util.framter. Accessed: 2025-01-08

[5] https://github.com/raspberrypi/Raspberry-Pi-OS-64bit/issues/168. Accessed: 2025-01-08

[6] https://github.com/ultralytics/yolov5/blob/master/requirements.txt. Accessed: 2025-01-08

Torch have stopped releasing official updates for 32-bit ARM devices meaning it was incompatible with the Raspberry Pi [7]. Older versions of PyTorch officially released for 32-bit ARM devices were too outdated for YOLOv5 to use. Therefore, an externally compiled version of PyTorch that runs 32-bit ARM devices was sourced and used [8].

The Intel Neural Compute Stick 2 (NCS2) Vision Processing Unit (VPU) (see Figure 2.27 H) contains a MYRIAD processor that is optimised to run machine vision problems at lower power. The NCS2 ran from the Raspberry Pi using the Raspberry Pi High Quality Camera. To run the NCS2 the OpenVINO Toolkit is required [9]. The OpenVINO Toolkit consists of the OpenVINO Development Tools and OpenVINO Runtime. The Development Tools can be used to convert models from PyTorch format to ONNX format and then to OpenVINO format [10]. If the model was created using a local version of YOLOv5 then the Development Toolkit would be required. However, using the Ultralytics Hub web application, the conversion to OpenVINO format was completed automatically online. OpenVINO Runtime is required before a model in OpenVINO format can run [11]. Installing OpenVINO Runtime on the Raspberry Pi can be a difficult process as some files will need to be altered to fix errors during the installation process. OpenVINO will only work on Intel platforms, therefore as the Raspberry Pi does not contain an Intel CPU, it will not run unless the NCS2 is connected [12]. Inside one of YOLOv5's setup files a line will need to be changed so that models in OpenVINO format will run on the NCS2 and not the CPU [13]. VPUs are designed to run machine vision algorithms and are optimised for performance per watt.

[7] https://discuss.pytorch.org/t/pytorch-on-a-raberry-pi-4-32-bit-os/138771. Accessed: 2025-01-08

[8] https://github.com/KumaTea/pytorch-arm. Accessed: 2025-01-08

[9] https://www.intel.com/content/www/us/en/developer/tools/openvino-toolkit/overview.html. Accessed: 2025-01-08

[10] https://docs.openvino.ai/nightly/openvino_docs_install_guides_install_dev_tools.html. Accessed: 2025-01-08

[11] https://docs.openvino.ai/nightly/openvino_docs_install_guides_installing_openvino_raspbian.htmldoxid-openvino-docs-install-guides-installing-openvino-raspbian. Accessed: 2025-01-08

[12] https://docs.openvino.ai/latest/openvino_docs_OV_UG_supported_plugins_Supported_Devices.html. Accessed: 2025-01-08

[13] https://github.com/ultralytics/yolov5/issues/8154. Accessed: 2025-01-08

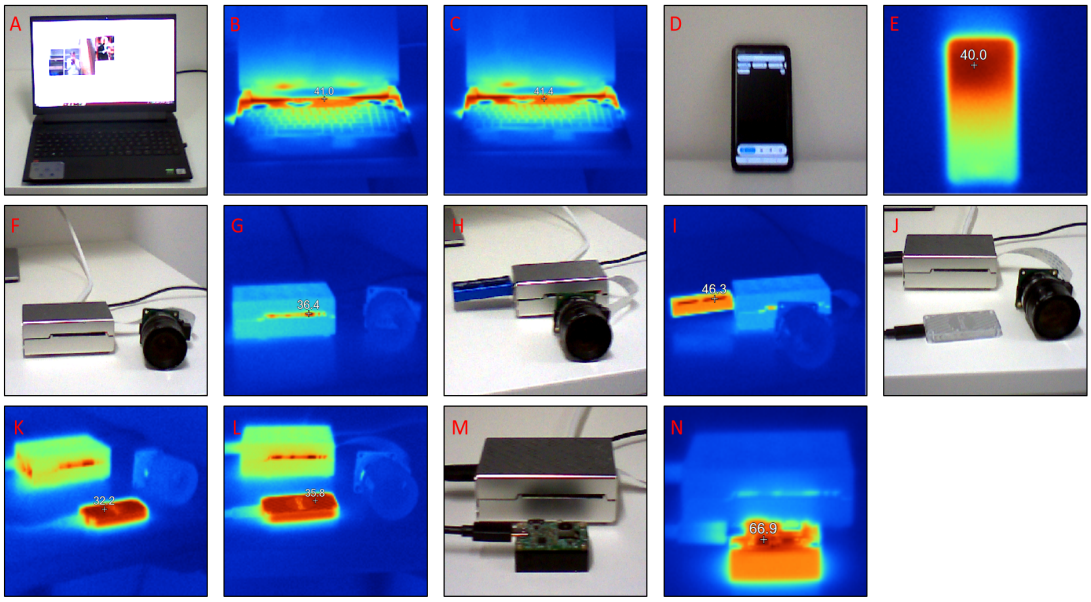

Figure 2.27: (A) A plugged-in laptop with an Intel Core i7-10870H CPU @ 2.20GHz CPU and a NVIDIA GeForce RTX 3060 Laptop GPU, (B) Thermal image of the plugged-in laptop running the model on its CPU, (C) Thermal image of the plugged-in laptop running the model on its GPU, (D) A Samsung Galaxy S20 FE 5G mobile phone, (E) Thermal image the mobile phone running the model, (F) Raspberry Pi 4 8GB RAM Model B CPU (32-bit Raspbian OS) with a Raspberry Pi High Quality Camera, (G) Thermal image the Raspberry Pi 4 Model B running the model, (H) Intel Neural Compute Stick 2 VPU running on the Raspberry Pi, (I) Thermal image the Intel Neural Compute Stick 2 running the model, (J) Coral USB Accelerator TPU running on the Raspberry Pi, (K) Thermal image of the Coral USB Accelerator at standard operating frequency running the model, (L) Thermal image the Coral USB Accelerator at max operating frequency running the model, (M) A Luxonis Oak-1-PCBA running on the Raspberry Pi, (N) Thermal image the Luxonis Oak-1-PCBA running the model.

The Coral USB Accelerator Tensor Processing Unit (TPU) ran from the Raspberry Pi using the Raspberry Pi High Quality Camera (see Figure 2.27 J). To run the Coral USB Accelerator the Edge TPU Runtime is required. There are two versions of the Edge TPU Runtime which run at either standard or maximum operating frequency. The Edge TPU Runtime running at maximum operating frequency should provide better performance but at the cost of higher energy usage and additional heat. TPUs are designed for use on with CNNs.

The Luxonis Oak-1-PCBA runs from the Raspberry Pi but not using the Raspberry Pi High Quality Camera (see Figure 2.27 M). The Oak-1 has its own camera attached to the device that allows extremely fast transmission of data from the camera to the device's Robotics Vision Core 2 removing the Raspberry Pi's CPU out of the process. The Oak-1 requires models in Blob File format to operate but neither the online Ultralytics Hub web application or local installation of YOLOv5 have a method to export models in Blob Files format, as of writing. Luxonis have provided a tool to allow the conversion of YOLO models from PyTorch to Blob Files [14]. The input shape varies between YOLOv5 models, but with the trained YOLOv5n6 model used to detect *Vespula vulgaris* the input shape was 448. The Oak-1 runs using DepthAI, with a program used to run YOLOv5 models in Blob File format [15]. This program was used instead of the usual detection script. It also has an inbuilt framerate counter.

The average framerate, power used and temperature of each device were recorded. Power consumption of the Laptop and Mobile phone were recorded using external applications on the respective devices; power consumption for the Raspberry Pi was displayed by a USB Multimeter. As the application used to record power consumption of the laptop only displayed wattage, the current and voltage are not available. The temperature of every device was measured using a Fluke Ti25 Thermal Camera. While recording the temperature each device ran the detection model until there was no change in temperature (accurate to 0.1°C), see Figure 2.27 B, C, E, G, I, K, L, and N.

[14] http://tools.luxonis.com/. Accessed: 2025-01-08

[15] https://github.com/luxonis/depthai-experiments/tree/master/gen2-yolo/device-decoding. Accessed: 2025-01-08

Table 2.5: File type, file extension, framerate, power consumption and temperature of each device while running the detection model

	Computer CPU	Computer GPU	Mobile Phone	Raspberry Pi	NCS2	Coral Std.	Coral Max	Oak-1
File Type	PyTorch	PyTorch	Android	PyTorch	OpenVINO	EdgeTPU	EdgeTPU	BlobFile
File Extension	.pt	.pt	.tflite	.pt	.bin,.xml,.mapping	.tflite	.tflite	.blob,.json
Framerate(Fps)	13.11	84.05	59	0.2	6.25	3.18	3.32	60.3
Potential(V)	N/A	N/A	3.808	5.06	4.94	4.95	4.96	4.88
Current(A)	N/A	N/A	1.919	0.75	0.312	0.19	0.21	0.62
Power(W)	58	16.5	7.31	3.8	1.54	0.94	1.04	3.03
Frames Per Watt	58	16.5	7.31	3.8	1.54	0.94	1.04	3.03
Temperature(°C)	41	41.4	40	36.4	46.3	32.2	35.8	66.9

The results show that in terms of framerate the Laptop GPU has the highest performance, while the Raspberry Pi 4 Model B CPU has the lowest performance (see Table 2.5, Figure 2.28). The Laptop CPU has a lower framerate performance but a higher power consumption than the Laptop GPU. This is most likely due to architecture differences between the CPU and GPU. The results also show that both the NCS2 and Oak-1 (which both utilise a VPU) outperform the Coral USB Accelerator TPU in framerate, but have a higher power consumption. Depending on whether a user finds framerate, power consumption or temperature more important will greatly impact their decision on choosing a device for live detection.

Framerate is essential to consider when choosing a device. Slow pests such as leatherjackets (crane fly/daddy long-legs larvae) *Nephrotoma appendiculata* would still be detected even in extremely low framerates, as they would remain in front of the camera for a longer duration. In contrast, fast pests such as the cabbage white butterfly *Pieris rapae* would require a detection algorithm running at a higher framerate to detect them.

It is necessary to examine power consumption when planning to leave a camera trap in operation for long periods of time. There are three methods of powering a camera trap: mains electricity, battery and a renewable energy source.

The use of mains electricity requires the camera trap to be near a socket connected to the mains. The use of mains electricity will remove the need to consider power consumption but it may impact the experiment as areas with a socket connected to the mains may not be an optimal place for detecting certain species of insects.

The use of a battery allows camera traps to be relocated anywhere but the power consumption needs to be considered. For example, a 12V 71Ah (255,600As) car

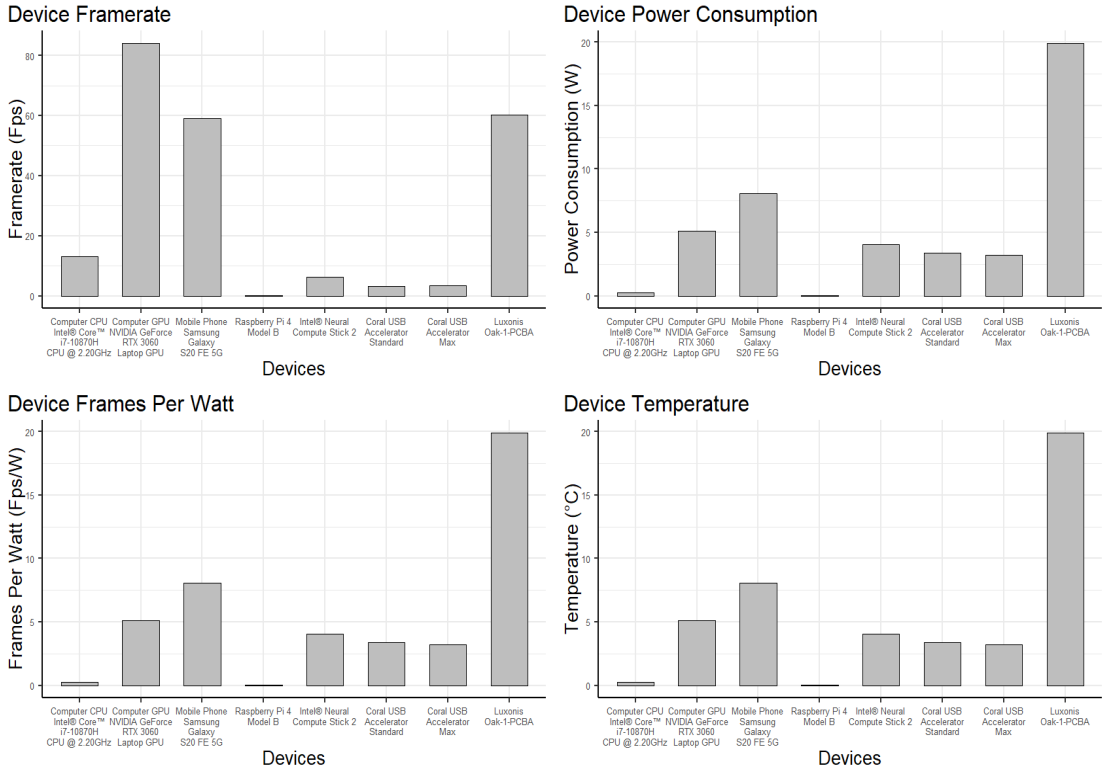

Figure 2.28: Device framerate, power consumption, framerate per watt and temperature.

battery would contain 3,067,200J of energy. The Computer CPU would be able to run on this battery for around 14.7 hours while the Coral USB Accelerator running at standard operating frequency would run for around 37.8 days.

Alternatively the use of renewable energy can provide a constant supply of power to a camera trap. Solar panels are an example of a renewable energy source that is portable. The most common silicone solar panels with an area of $1m^2$ produce 150W of power on a clear sunny day and on average 130kWh or 468MJ a year. Averaging the power generated from a $1m^2$ solar panel regardless of the time of day or year, 15W should be produced. Therefore, the Computer CPU would require a solar panel with an area greater than $3.87m^2$ to supply sufficient power, while the Coral USB Accelerator running at standard operating frequency would only require a solar panel with an area greater than $0.06m^2$ (e.g. larger than a

20cm × 30cm panel). Using solar panels like this would still require a battery so the camera trap could be powered at night and during Winter.

The temperature produced by a device is also an essential factor to consider. Some heat is useful while using a camera trap in the field, as it removes condensation from camera lenses which is essential during colder months of the year. However, heat may affect the environment and therefore influence which insects approach the camera trap; this issue could be amplified if noise-producing devices such as fans are required for cooling. The use of cooling will also need to be accounted for within the energy budget if the device is not connected to mains electricity.

The equations used to calculate the energy requirements are

$$\begin{aligned} As &= 3,600Ah \\ J &= VAs = Ws \\ P &= VA \\ Wh &= 3,600Ws \end{aligned}$$

where J is energy in Joules, V is potential difference (voltage), A is current (amperage), As is continuous current flowing for a second (Ampere-Seconds), Ah is continuous current flowing for an hour (Ampere-Hours), P is power in Watts (W), Ws is continuous power used for a second (Watt-Second), and Wh is continuous power used for an hour (Watt-Hour).

Overall, the best suited device for detecting animals in the field is a device that is relatively small, can run live detection models at high framerates, requires low power and does not produce much heat. However, the device that is most optimal for a task is very dependent on the environment and species that is to be observed. These points align directly with the implementation of a monitoring system for conservation biology purposes, specifically the target of Caughley's management action, by providing insights that can be applied when designing those systems. These points are also not restricted to entomology studies and are transferable to other fields where animal monitoring is required.

2.2.6 Final considerations

We have provided an overview of machine vision methods to identify and localise insects in the laboratory or in the field. We discussed the use of different machine learning and computer vision techniques, as well as software and hardware. There is a range of available methods, and machine vision as a field of study is constantly evolving. Therefore, we aimed to present the reader with a handful of options currently available for identification and localisation tasks, but we highlight that the choice of which methodology to use is dependent on the question and species of interest. We identified two main limitations of this work. The first relates to the scarcity and imbalance of the data presented in the examples. However, our findings demonstrate the feasibility of these methods through these proof-of-concept examples. In future work, additional data can be gathered, class imbalance addressed, and other metrics that better account for class imbalance, such as F1-score and confusion matrices, can be presented. The second limitation relates to the first example, introducing a dataset of medically and forensically important flies that pose specific challenges in the given situation. However, the example demonstrates the feasibility of the proposed methods, and, as future work, more focus can be given to additional examples outside a Museum scenario, where pins are available, and the scene is explicitly designed for this context.

2.3 Towards species' classification of the *Anastrepha pseudoparallela* group

In this final application, the contribution to the studied management action involves proposing a new approach to classify insect species that present common challenges in the entomological literature: the group's rare species characteristics, which result in imbalanced, scarce data. The *Anastrepha* species from the *pseudoparallela* group present a clear example of these challenges. The focus of this chapter is to classify five species in this group and to recommend a new approach by exploring new features, machine learning methods, and techniques to address data scarcity and imbalance. The identification of these species utilises the morphological characteristics of the specimens' wings and aculeus tips. Considering the importance of identifying these species of this group, this application

contributes to automating the classification of the species *Anastrepha chiclayae* Greene, *Anastrepha consobrina* (Loew), *Anastrepha curitibana* Araújo, Norrbom & Savaris, *Anastrepha curitis* Stone, and *Anastrepha pseudoparallela* (Loew) using deep Learning methods and designing new features based on wing structures. Automating classification for this group is challenging due to limited data availability and class imbalance in the dataset. We explored transfer learning approaches and proposed a new set of features tailored to each wing's structure. We used the dual annealing algorithm to optimise the hyperparameters of Deep Neural Networks, Random Forests, Decision Trees, and Support Vector Machines, combined with autoencoders and SMOTE to address class imbalance. We tested our approach on a dataset of 127 high-quality images from 5 species. We used three-fold cross-validation for training, tuning, and testing, with six permutations to assess the performance of the learning algorithms. Our findings demonstrate that our novel approach, which combines feature extraction and machine learning techniques, can improve the species classification accuracy for rare *Anastrepha pseudoparallela* group specimens, with the SMOTE and Random Forests algorithms leading to the average performance of 0.72 in terms of the mean of the individual accuracies considering all species. Our results are promising for classifying rare species using small, imbalanced datasets.

2.3.1 Introduction

The classification and identification of insects is fundamental for comprehending global biodiversity, given that this group represents an expressive percentage of the available biodiversity today [Gaston, 1991, Sankarganesh, 2017, Wagner et al., 2021, Chowdhury et al., 2023, Vaz et al., 2023, Hailay Gebremariam, 2024]. A clear picture of the global biodiversity yields a better understanding of the ecological services that new species can potentially provide and enlightens conservation and management practices [Thrupp, 2004, Gamfeldt et al., 2008, Uchida et al., 2021, Upreti, 2023, Riva et al., 2024]. Entomologists play a fundamental role in identifying taxa and understanding the evolutionary, ecological and functional relationships among these taxa [van Noort, 2024]. The study of biodiversity is

heavily reliant upon correct classification of animal specimens. The identification can be performed in various ways, using morphometric features [El-Ahmady et al., 2024, Rodrigues et al., 2024, Laojun et al., 2024], DNA-based identification, such as the barcoding method [Chua et al., 2023, Srivathsan et al., 2024], acoustic features [Chen et al., 2014, Phung et al., 2017, Hibino et al., 2021, He et al., 2024, Branding et al., 2024] and other methods [Raffini et al., 2020, Høye et al., 2021b, Van Klink et al., 2022, Karbstein et al., 2024]. In this context, there is an opportunity to apply different methods targeting insect automatic monitoring [Van Klink et al., 2022, Høye et al., 2021b, Palma et al., 2023b].

Various researchers have analysed the efficiency of machine learning, deep learning, computer vision and a combination of these methods when applied to insect classification, especially for pest species [Passias et al., 2024, Assiri et al., 2024]. The results obtained from these studies have been positively affecting Integrated Pest Management (IPM) protocols by providing data-driven decision-making systems [Gao et al., 2024, Moonis and Singh, 2024, Amrani et al., 2024]. However, challenges have been reported related to the use of these techniques to automate insect monitoring, such as the high similarities of insect species, obtaining large and high-quality annotated datasets, and the presence of imbalanced classes due to lack of homogeneity in the same groups of insects [Nawoya et al., 2024, PISE and PATIL, 2024].

In other research areas, several authors have proposed adaptations of deep learning methods to accommodate the low data availability, such as changes in Convolutional Neural Networks (CNN) architectures by including regularisation techniques, such as drop out and L_1 regularization [Brigato and Iocchi, 2021, Koppe et al., 2021]. Transfer learning is commonly reported as an alternative solution to increase the performance of deep learning models for image classification [Barbero-Aparicio et al., 2024, Rachman et al., 2024]. Also, several data-augmented approaches have been implemented to increase the number of samples, such as Synthetic Minority Over-sampling Technique (SMOTE), autoencoders (AEs), variational autoencoders (VAEs), generative adversarial networks (GANs) and adaptations of deep generative modelling methods using additional layers such as convolution, max pooling and others [van Tilborg et al., 2024, Mumuni et al., 2024].

In entomology, researchers have recently proposed the use of augmentation techniques to increase the number of samples for species of fruit fly [Shen et al., 2024, Medina-Ramos et al., 2024, Zhang et al., 2024]. Also, the use of transfer learning is a common practice for the classification of fruit flies [Leonardo et al., 2018b, Martins et al., 2019b, Gosaye and Moloo, 2022, Slim et al., 2023, Molina-Rotger et al., 2023].

For imbalanced datasets, several techniques have been implemented to deal with this issue, such as SMOTE, undersampling, oversampling, and combining transfer learning and active sampling [Yuan et al., 2023, Liu et al., 2023, Wongvorachan et al., 2023, Rezvani and Wang, 2023]. Specifically, rotation and change in the background are commonly reported in the literature as alternatives for data augmentation. In entomology, researchers have implemented these techniques for different taxa, belonging to Diptera, Coleoptera, Hymenoptera, and Lepidoptera [Bjerge et al., 2023, PISE and PATIL, 2024, Doan, 2023]. Moreover, the use of generative deep learning methods, such as autoencoders, variational autoencoders, and other variations, has been commonly reported in the literature as an alternative to generating new data [Cabrera and Villanueva, 2021, Klasen et al., 2022, Borowiec et al., 2022, Nitin et al., 2023, Phong et al., 2024]. These techniques have shown promising results, and more researchers have started implementing them in their automation frameworks [Al-Shahari et al., 2024, Khan et al., 2024].

One example of the combination of imbalanced and low data availability challenges is the *Anastrepha pseudoparallela* group (Diptera: Tephritidae) [Araújo et al., 2024]. Identifying *Anastrepha* species from this group is problematic due to morphological similarities among species and a broad geographic variation [Araújo et al., 2024]. This group comprises 31 species of fruit flies and includes pests that affect passion fruit crops [Norrbom et al., 1999, Malavasi and Zucchi, 2000]. Some species of this group are easily encountered, and others are rare in the field, generating a class imbalance problem for this group.

Therefore, we propose a new framework to identify species of this group, specifically, *Anastrepha chiclayae* Greene, *Anastrepha consobrina* (Loew), *Anastrepha*

curitibana Araújo, Norrbom & Savaris, *Anastrepha curitis* Stone, and *Anastrepha pseudoparallela* (Loew) due to the agronomic and ecological importance of this group combined with the challenge posted at the quantitative methods for automating the monitoring of these species [Araújo et al., 2024]. This this chapter aims to propose a new framework combining the wing's morphological features collection with machine learning algorithms and data augmentation techniques to automate the identification of species of the *Anastrepha pseudoparallela* group. We explored the use of different learning algorithms, preprocessing, feature extraction methods and augmentation techniques to provide a novel approach to identifying species of this group.

The rest of this this chapter is organised as follows. In Section 2.3.3, we introduce the proposed approach for extracting and collecting features from the images. In Section 2.3.3.1, we present the main methods used for data augmentation. In section 2.3.3.2, we introduce the learning algorithms used to classify the five species of the *pseudoparallela* group and the selected approach for hyperparameter tuning. Finally, in Section 2.3.4, 2.3.5, and 2.3.6, we, respectively, present the experimental results, discuss our findings and present our overall conclusions from this application.

2.3.2 Methods

2.3.3 Image processing and feature extraction

For image acquisition, the right wing of 127 females from five species of the *Anastrepha pseudoparallela* group was first detached from the thorax and submerged in Celossolve (C4H10O2) for 3-5 days. Then, it was mounted on permanent slides containing Euparal® and dried for seven days in a laboratory oven at 25ºC. The wing pattern of each specimen was photographed with a Leica DFC 450 camera coupled with a Leica M205 stereomicroscopic. The final dataset is composed by 21, 16, 72, 11 and 7 images of respectively *A. curitibana*, *A. pseudoparallela*, *A. chiclayae*, *A. curitis*, and *A. consobrina* species. Initially, we first removed the white background using RGB thresholding and the Canva application to remove any noise in the wing's images. Then, we preprocessed these images to create three distinct datasets composed by coloured, gray-scale, and histogram-

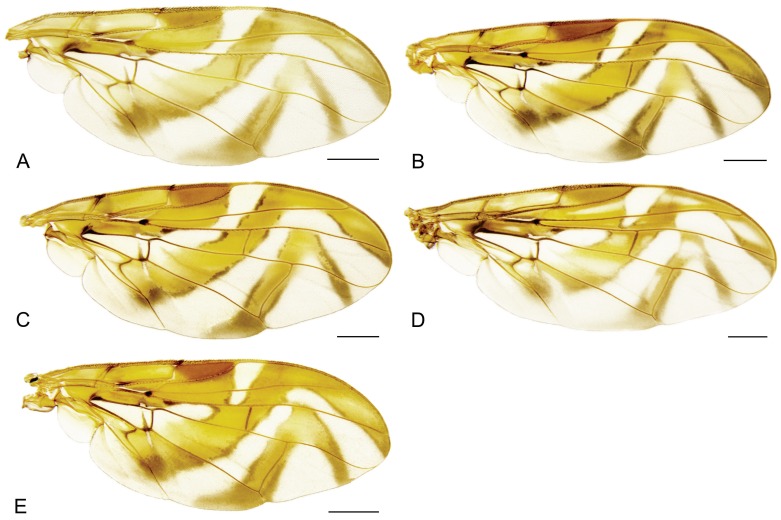

Figure 2.29: Wings of (A) *Anastrepha chiclayae*; (B) *Anastrepha consobrina*; (C) *Anastrepha curitibana*; (D) *Anastrepha curitis*; (E) *Anastrepha pseudoparallela*. Scale bars = 1.00 mm

equalizedimages [Palma et al., 2022]. We used the OpenCV package [Bradski, 2000] in most image processing techniques applied to create these datasets. Finally, we combined the Convolutional Neural Network architecture VGG16 trained with the imagenet dataset [Simonyan and Zisserman, 2014] and Principal Component Analysis (PCA) for feature extraction from each wing dataset, as explained below [Kaur and Singh, 2024, Singh et al., 2024a].

We used this approach to avoid adverse effects on the learning algorithms' performance due to the high dimension of the feature vector produced by the VGG16 architecture. This approach allows us to explore classical machine learning and deep learning algorithms to classify the species of the *Anastrepha pseudoparallela* group. The coloured wing images have dimensions of 2560×1920 pixels, which would produce a VGG16 feature vector of $2,457,600$ elements. Therefore, we first reduced the image size to 256×192 pixels, producing VGG16 feature vectors of $24,576$ elements. Finally, we applied PCA to reduce dimensionality and mitigate potential performance degradation in classical machine learning algorithms. The number of principal components retaining 95% of the variability for the coloured, gray-scale and histogram-equalised datasets were 87, 89, and 87, respectively.

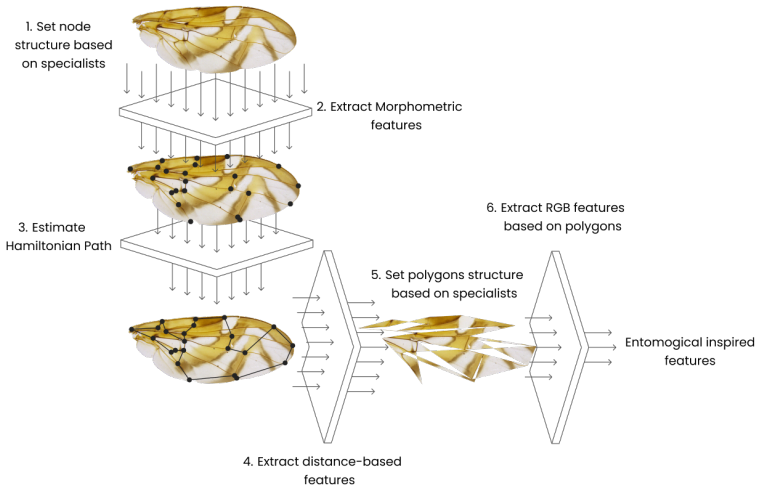

Figure 2.30: Diagram illustrating the proposed feature extraction using morphometric and RGB data based on the proposed approach using distance and colour-based features with, respectively, the shortest HC and polygon structures based on specialist input.

We systematically collected classical morphometric features, including wing length and height measurements. Additionally, we identified features based on the structural composition of the wings. This process involved analysing the nodes and polygons constituting critical cells within the wing structure. We selected the distances between each pair of nodes and the total distance of all pairs as features. In addition, we used the polygons produced by each important cell that are taxonomic relevant to the species identification and collected the average, 2.5% and 97.5% percentile of RGB from all polygons designed for each image [Palma et al., 2023b].

To define the pairs of nodes, we assumed that each wing's node is a vertex of an undirected graph and identified the shortest Hamiltonian Cycle (HC) for the wing's structure. The shortest HC is a closed loop in a graph that visits each vertex exactly once and returns to the starting vertex [Lozin, 2024]. Finding the shortest HC in a graph is a well-known computer science problem named the Travelling Salesman Problem (TSP) [Carmesin et al., 2023]. In the context of our work, each node represents a key structural point of the wing. This way, we ensure that only the most relevant geometric relationships between nodes are captured while

avoiding the inclusion of redundant or less informative connections. Overall, we propose using the pair-wise distances of the selected nodes belonging to the TSP's solution combined with RGB features of the selected polygons formed by regions of interest on the wings and the wing's length and height. Figure 2.30 illustrates the proposed approach to collect these features for a given shortest HC identified in a wing.

We used the Ant Colony Optimisation (ACO) algorithm to find a TSP solution for the wing's structure considering its common use in the literature to solve this problem [Dorigo and Gambardella, 1997, Stützle et al., 1999, Li and Gong, 2003]. The Ant Colony Optimisation (ACO) algorithm is a probabilistic approach inspired by the behaviour of ants seeking a path between their colony and a food source [Dorigo et al., 2006, Pedemonte et al., 2011, Mohan and Baskaran, 2012, Dokeroglu et al., 2019, Tang et al., 2021]. The pseudo code 1 introduces an overview of the mechaniques of this algorithm. We used 2000 ants, $\alpha = 1$, $\beta = 2$, $\rho = 0.5$, $Q = 100$, and 200 iterations.

2.3.3.1 Data augmentation

Considering the challenge of having solely 127 images belonging to five classes, we addressed this problem by implementing data augmentation to increase the number of species samples with fewer images to balance the dataset. We used two approaches to perform this task: the SMOTE (Synthetic Minority Over-sampling Technique) [Chawla et al., 2002, Fernández et al., 2018, Khan et al., 2024] and deep generative modelling based on an autoencoder architecture [Dong et al., 2018, Pratella et al., 2021] to increase the number of samples from the minority class. SMOTE is a data augmentation algorithm that generates synthetic minority class samples by interpolating between existing minority class instances in the feature space [Chawla et al., 2002, Fernández et al., 2018]. A classical autoencoder is a variation of an Artificial Neural Network (ANN) architecture, which contains an encoder, latent and decoder components, where the encoder and decoder contain ANN's layers and the same number of neurons. The latent component is an ANN's layer with fewer number of neurons, which is commonly used for encoding, dimensional reduction, and data generation [Sarroff and Casey, 2014, Thakkar et al.,

Algorithm 1 Wing Feature Extraction using ACO-based TSP

Require: Wing image I
Require: ACO parameters: N_{ants}, T, α, β, ρ, Q
Ensure: Feature vector F

Step 1: Initialize structures (I) Extract structural nodes N from wing image I; (II) Extract taxonomic cell polygons P from wing image I; (III) Initialize empty feature vector F
Step 2: Extract proposed morphometric features (I) Calculate wing length l as maximum distance between anterior-posterior nodes; (II) Calculate wing height h as maximum distance between dorsal-ventral nodes; (III) Add l, h to feature vector F
Step 3: ACO algorithm for TSP (I) Initialize pheromone matrix τ_{ij} by sampling values from a uniform distribution; (II) Calculate visibility matrix $\eta_{ij} = 1/d_{ij}$ where d_{ij} is distance between nodes i, j; (III) Initialize best solution S_{best} with $L_{best} = \infty$
for $t = 1$ to T **do**
    Construct N_{ants} solutions using probability $p_{ij}^k = \frac{(\tau_{ij})^\alpha (\eta_{ij})^\beta}{\sum_{l \in U_k} (\tau_{il})^\alpha (\eta_{il})^\beta}$
    Update S_{best} if better solution found
    Update pheromone trails: $\tau_{ij} = (1 - \rho)\tau_{ij} + \sum_k \Delta\tau_{ij}^k$
    where $\Delta\tau_{ij}^k = Q/L_k$ if edge (i, j) is in ant's path
end for
Step 4: Extract distances from best Hamiltonian path (I) Calculate sequential node distances from S_{best}; (II) Add node distances to feature vector F
Step 5: Extract colour features from polygons
for each polygon p in P **do**
    Calculate mean and percentile (2.5%, 97.5%) RGB values
    Add RGB statistics to feature vector F
end for
Return: Feature vector F

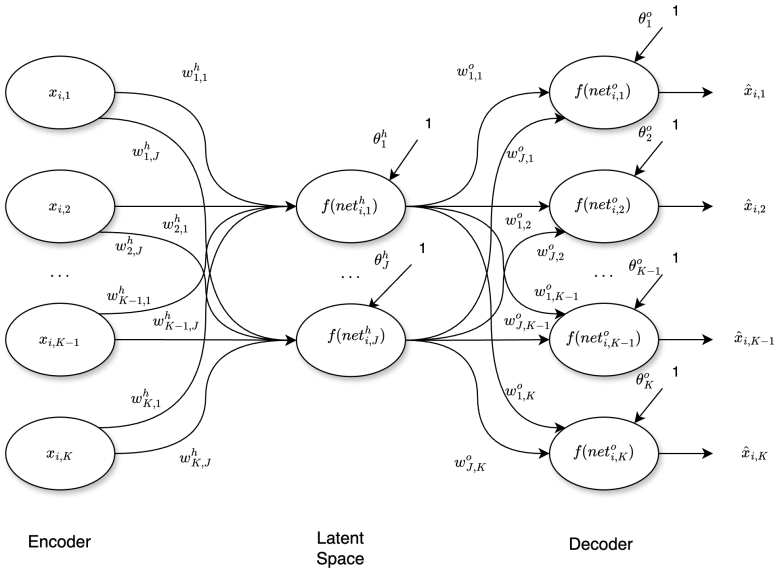

Figure 2.31: Diagram of a simple autoencoder containing one layer with J and K neurons for the latent space and encoder/decoder components.

2019, Mansouri and Lachiri, 2020, Yoon et al., 2022, Govender et al., 2024].

The final balanced datasets based on SMOTE and autoencoder contained 360 and 720 samples. The SMOTE algorithm contains fewer data points due to the algorithms' classical structure of increasing solely the minority classes [Chawla et al., 2002, Fernández et al., 2018], and we increased the number of samples of the minority classes to equal the majority class. We used the same approach for the autoencoder algorithm, and considering its flexibility [Jeong et al., 2022], we evenly included additional samples for all classes.

Figure 2.31 presents a standard autoencoder architecture containing one layer for the encoder, decoder and latent components, where $w^h_{k,j}$ and $w^o_{k,j}$ are the weights, θ^h_j and θ^o_1 are the biases for respectively, the encoder and decoder components. $net^h_{i,j}$ and $net^o_{i,K}$ represents linear combinations and f is an activation function. We used two layers for the encoder with n and 30 neurons, where n corresponds to the number of input features of a presented dataset. For the decoder, we also used two layers; the first performs batch normalisation operation, and the second contains n neurons. We used a *Leaky Rectified Linear Unit* (LeakyReLU) activation function,

and we included a custom loss function defined as the combination of normalised Euclidean distance and the correlation coefficient between the true and predicted values [Deng et al., 2017, Wang et al., 2018, Wu and Picek, 2020, Hu et al., 2023]. We selected the following loss function after testing a few options and evaluating the performance of the learning algorithms:

$$\mathcal{L} = \frac{||\mathbf{y}_{\text{true}} - \mathbf{y}_{\text{pred}}||_2}{\sqrt{\sum \left(\max(\mathbf{y}_{\text{true}}) - \min(\mathbf{y}_{\text{true}})\right)^2 + \epsilon}} - \rho(\mathbf{y}_{\text{true}}, \mathbf{y}_{\text{pred}}), \tag{2.1}$$

where $\rho(.)$ is the algebraic correlation between the true, $\mathbf{y}_{\text{true}}$, and predicted, $\mathbf{y}_{\text{pred}}$, output vectors, respectively. $||\mathbf{y}_{\text{true}} - \mathbf{y}_{\text{pred}}||_2$ represents the Euclidean distance and ϵ is a small constant to prevent division by zero. This loss function minimises the distance, thereby reducing the reconstruction error while simultaneously maximising the correlation between the true and predicted values to preserve the inherent relationships within the data. Finally, we used the *Adadelta* algorithm [Zhang et al., 2023a] to update the parameters of the autoencoder structure. We varied the autoencoder structure to explore the effects of a multi-headed attention mechanism[Vaswani, 2017] on the data augmentation process and, consequently, the learning algorithm's performance.

2.3.3.2 Validation approach and learning algorithms

Overall, considering the pre-processing techniques applied to the images, augmentation techniques and the feature extraction procedures, we applied different learning algorithms to 12 datasets:

- Three datasets containing VGG16-PCA features from coloured, gray-scale, and histogram equalised images with augmented data based on the autoencoder algorithm;
- Three datasets containing VGG16-PCA features from coloured, gray-scale, and histogram equalised images with augmented data based on the SMOTE algorithm;

- Three datasets containing VGG16-PCA features from coloured, gray-scale, and histogram equalised images with no data augmentation;
- Three datasets containing the proposed entomologically-inspired features from coloured, gray-scale, and histogram equalised images without data augmentation, as well as augmented data based on the SMOTE and the autoencoder algorithms;

For the validation procedure, we split each dataset into three folds by dividing the number of samples evenly for each fold. The first fold was used for training, the second for fine-tuning and the third for obtaining the test performance to observe the generalisation potential of each algorithm. We used six permutations of these folds and calculated the mean of individual accuracies for each permutation. These permutations relate to the use of each created fold. Since we have three folds, the total number of permutations is 6!. For each permutation, the data are independent, and a possible scenario of train, fine-tuning, and test is presented to evaluate the variance of the performance metrics. We ensure that we introduce only synthetic data not included in the testing fold for all permutations. We implemented four learning algorithms for classifying the *pseudoparallela* species, Random Forests (RF), Support Vector Machines (SVM) with polynomial kernel, Decision trees (DT) and Deep Neural Networks (DNNs). We used the Scikit-learn [Pedregosa et al., 2011] and TensorFlow [Abadi et al., 2015] packages.

During the optimisation of the hyperparameters of each algorithm, we use the dual annealing algorithm [Tsallis and Stariolo, 1996], which is a stochastic algorithm used for finding the global minimum of a given function. Briefly, the objective function used to optimise the parameters of each learning algorithm computes the negative value of the average of the individual accuracies per class obtained by the learning algorithms using a test dataset. Then, the Generalised Simulated Annealing method uses 200 visits on the objective function to find the optimum combination of parameters. Table 2.6 shows the hyperparameters optimised by the dual annealing algorithm, the boundaries used for each hyperparameter, and the learning algorithm's nature. Additionally, we optimised the image size used for the learning algorithms. We used Python for image processing,

machine learning and data augmentation methods and R for data visualisation. All code and datasets are available at https://github.com/GabrielRPalma/AnastrephaPsedoparallelaClassification to allow full reproducibility of this work.

Table 2.6: Attributes of the learning algorithms used for classifying species of the `pseudoparallela` group including the type of hyperparameter and boundaries used in the dual annealing algorithm during the optimisation process.

	Attributes		
Learning algorithms	Hyperparameters	Description	Boundaries
Decision Trees (DT)	j, s, c	j is the maximum depth of the created trees, c is the minimum sample leafs and c is the minimum sample split	$j = \{2, \ldots, 50\}$, $s = \{2, \ldots, 50\}$ and $c = \{2, \ldots, 50\}$
Deep Neural Network (DNN)	l, u, d and α	l is the number of layers, u is the total number of neurons of the neural network, d indicates dropout is implemented in the and α is the negative slope used for the *Leaky Rectified Linear Unit* (LeakyReLU) activation function in the DNN architecture	$l = \{1, \ldots, 7\}$, $d = \{0, 1\}$, $u = \{10, \ldots, 50\}$ and $\alpha = \{0, \ldots, 1\}$
Support Vector Machines (SVM)	p	The polynomial degree of the applied kernel used in the method	$p = \{1, \ldots, 20\}$
Random Forest (RF)	b, j, s	b is an indicator to apply bootstrapping, j is the maximum depth of the created trees and s is the minimum sample split	$b = \{0, 1\}$, $j = \{2, \ldots, 50\}$ and $s = \{2, \ldots, 50\}$

2.3.4 Experimental results

Figure 2.32 presents the performance obtained by the Deep Neural Networks, Support Vector Machines, Decision Trees, and Random Forests algorithms to classify species of the species *Anastrepha pseudoparallela* group based on the proposed and VGG16-PCA features. Overall, considering all machine learning and data augmentation algorithms, we obtained an average and standard deviation (sd) of the mean of individual accuracies for the Travel Salesman Problem (TSP)-based and VGG16-PCA features of, respectively, 0.51 (0.19) and 0.35 (0.22). Also, considering the pre processing the images, we obtained for the coloured, gray scale and

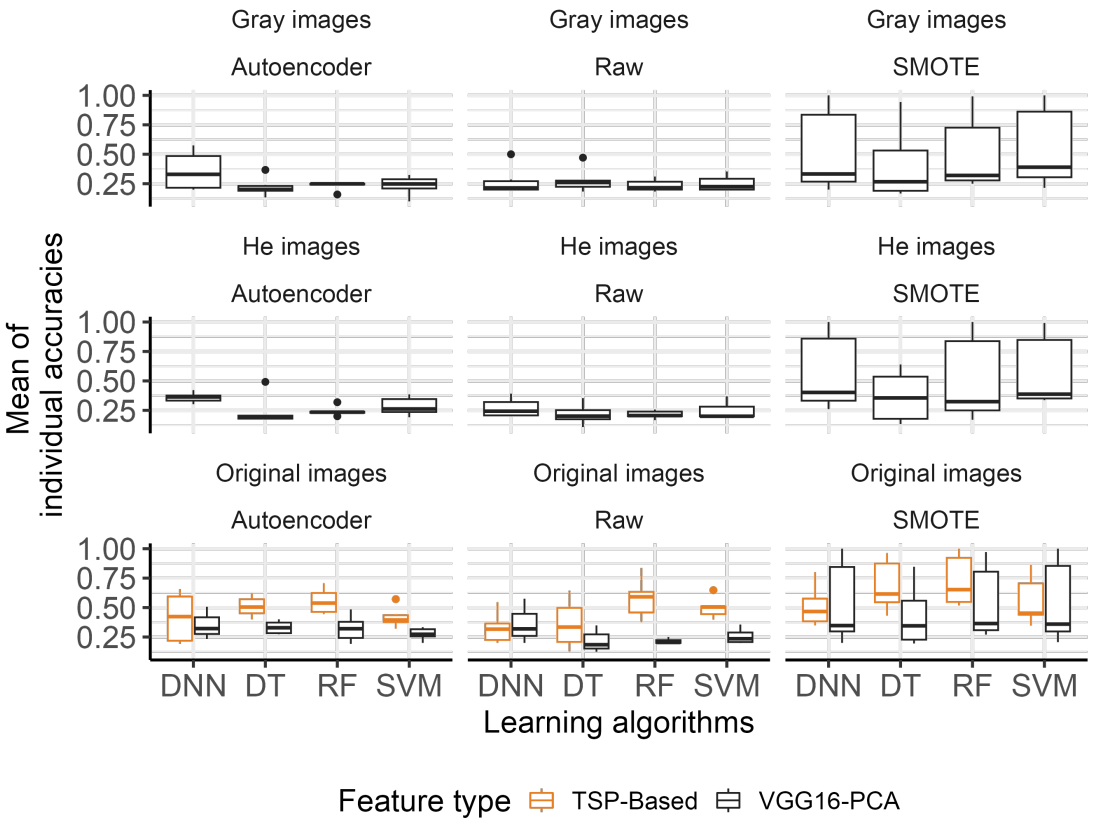

Figure 2.32: Mean of individual accuracies of four machine learning methods applied to the classification of *A. consobrina*, *A. curitis*, *A. chiclayae* and *A. curitibana* based on images of their wings using. All measures reported in this table were obtained in the test set.

histogram equalised images an average performance of, respectively, 0.44 (0.22), 0.34 (0.23), and 0.34 (0.22). Moreover, we obtained an average and standard deviation (sd) of the mean of individual accuracies for SMOTE, autoencoder, and raw datasets of, respectively, 0.53 (0.29), 0.33 (0.13), and 0.30 (0.14). Random forests, Deep Neural Networks, Decision Trees, and Support Vector Machines algorithms obtained an average performance of 0.40 (0.25), 0.40 (0.22), 0.35 (0.21), and 0.39 (0.23).

For the classification of *A. chiclayae* considering all machine learning and data augmentation algorithms, we obtained an average of the mean of individual ac-

curacies for the TSP-based and VGG16-PCA features of, respectively, 0.78 (0.24) and 0.75 (0.29). Also, considering the pre processing the images, we obtained for the coloured, gray scale and histogram equalised images an average performance of, respectively, 0.77 (0.26), 0.74 (0.32), and 0.75 (0.27). Moreover, we obtained an average of the mean of individual accuracies for SMOTE, autoencoder, and raw datasets of, respectively, 0.72 (0.22), 0.70 (0.31), and 0.84 (0.21). In addition, Random Forests, Deep Neural Networks, decision trees, and Support Vector Machines algorithms obtained an average performance of 0.81 (0.23), 0.87 (0.19), 0.56 (0.30), and 0.78 (0.27).

For the classification of *A. consobrina* considering all machine learning and data augmentation algorithms, we obtained accuracies for the TSP-based and VGG16-PCA features of, respectively, 0.32 (0.41) and 0.28 (0.36). Also, considering the pre processing the images, we obtained for the coloured, gray scale and histogram equalised images an average performance of, respectively, 0.32 (0.38), 0.27 0.36), and 0.27 (0.36). Moreover, we obtained accuracies for SMOTE, autoencoder, and raw datasets of, respectively, 0.52 (0.41), 0.23 (0.32), and 0.13 (0.26). In addition, Random Forests, deep neural network, decision tree and Support Vector Machines algorithms obtained performances of respectively 0.296 (0.38), 0.31 (0.39), 0.28 (0.36), and 0.29 (0.37)

For the classification of *A. curitis* considering all machine learning and data augmentation algorithms, we obtained accuracies for the TSP-based and VGG16-PCA features of, respectively, 0.70 (0.36) and 0.24 (0.34). Also, considering the pre processing the images, we obtained for the coloured, gray scale and histogram equalised images an average performance of, respectively, 0.48 (0.42), 0.25 (0.34), and 0.23 (0.33). Moreover, we obtained accuracies for SMOTE, autoencoder, and raw datasets of, respectively, 0.67(0.40), 0.58(0.37), 0.39(0.38), and 0.47(0.41). In addition, Random Forests, deep neural network, Decision Trees, and Support Vector Machines algorithms obtained an average performance of 0.36 (0.41), 0.34 (0.41), 0.38 (0.39), and 0.35 (0.39).

For the classification of *A. pseudoparallela* considering all machine learning and data augmentation algorithms, we obtained accuracies for the TSP-based and

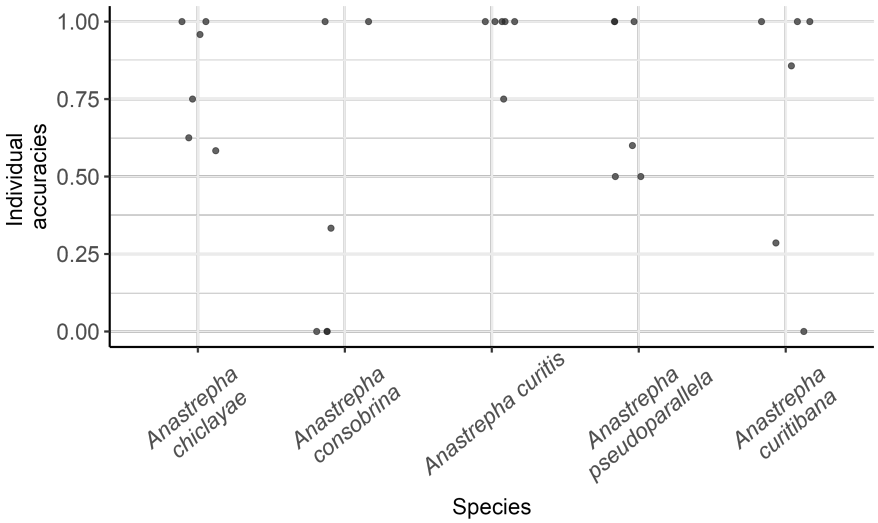

Figure 2.33: Random Forests' individual accuracies of each species of the *Anastrepha* group based on the features collected from the coloured images of their wings using the SMOTE algorithm for data augmentation. All measures reported in this table were obtained in the test set.

VGG16-PCA features of, respectively, 0.43 (0.36) and 0.20 (0.34). Also, considering the pre processing the images, we obtained for the coloured, gray scale and histogram equalised images an average performance of, respectively, 0.32 (0.36), 0.20 (0.34), and 0.20 (0.35). Moreover, we obtained an average of the mean of individual accuracies for SMOTE, autoencoder, and raw datasets of, respectively, 0.47 (0.42), 0.212 (0.30), and 0.10 (0.23). In addition, Random Forests, deep neural network, Decision Trees, and Support Vector Machine algorithms obtained an average performance of 0.25 (0.35), 0.19 (0.34), 0.30 (0.36), and 0.29 (0.38).

For the classification of *A. curitibana* considering all machine learning and data augmentation algorithms, we obtained accuracies for the TSP-based and VGG16-PCA features of, respectively, 0.32(0.36) and 0.25 (0.32). Also, considering the pre processing the images, we obtained for the coloured, gray scale and histogram equalised images an average performance of, respectively, 0.30 (0.35), 0.24 (0.34), and 0.25 (0.29). Moreover, we obtained accuracies for SMOTE, autoencoder, and raw datasets of, respectively, 0.41 (0.40), 0.22 (0.27), and 0.19 (0.28). In addition, Random Forests, deep neural network, Decision Trees, and Support Vector Ma-

chines algorithms obtained performances of 0.30 (0.35), 0.29 (0.38), 0.25 (0.30), and 0.24 (0.30).

Figure 2.33 shows the individual accuracies per species of the studied group, where the Random Forests algorithm was employed using the coloured images for feature collection and the SMOTE algorithm. This combination presents the highest average performance we could obtain for the classification of this group, yielding an average performance of individual accuracies of 0.72 (0.22). The average individual accuracies performance for *A. chiclayae*, *A. consobrina*, *A. curitis*, *A. pseudoparallela*, and *A. curitibana* were, respectively, 0.82 (0.19), 0.39 (0.49), 0.96 (0.10), 0.77 (0.26), and 0.69 (0.44). We noticed a high level of variability in monthly performance for *A. consobrina*, as solely 7 images were available.

2.3.5 Discussion

We proposed a new set of features based on the shortest Hamiltonian Cycle extracted in the wing's structure and compare them with features obtained by the VGG16-PCA architecture to classify species of the *A. pseudoparallela* group with various learning algorithms and data augmentation techniques. Our results showed that using the Random Forests algorithm trained with the proposed features collected from the coloured images and augmented using the SMOTE algorithm yielded higher average performance than the other presented combinations. The poor performance of these approaches demonstrates how challenging this problem is and highlights potential avenues for improving this work. Our results agreed with the broader literature on ensemble learning applied to imbalanced data [Rezvani and Wang, 2023, Khan et al., 2024]. Our results provided a novel contribution to the *pseudoparallela* group by introducing the proposed approach to classify closely related species given the challenges of imbalanced and low data availability.

Moreover, our results indicated that the use of the proposed features collected by the solving TSP based on the wing's nodes and collecting specific distances based on the resulting shortest HC provide higher average performance compared to the use of the VGG16-PCA applied to the high-resolution images of the wings. Therefore, it presented a promising alternative for researchers looking for data dimensionality reduction and classification of winged insect species. Moreover,

using coloured images instead of image transformations (gray scale and histogram equalised) showed better overall performance, indicating no need for modifying the RGB structure of the image even though for the human eyes, the histogram equalisation would emphasise the node structure of the wings. Finally, the SMOTE algorithm also increases the algorithm's overall performance in classifying this group.

A. consobrina specimens presented the most challenge for the learning algorithms, even for the best approach proposed in this this chapter. The lack of available images can explain this, and the increase in average performance by the SMOTE algorithm highlighted this possible explanation. Also, the classification of the studied group showed to be more challenging compared to other species of *Anastrepha*, such as *A. fraterculus* (Wiedemann), *A. obliqua* (Macquart), and *A. sororcula* Zucchi [Perre et al., 2016, Leonardo et al., 2018b] and our work highlights the importance of proposing new techniques for automatically identifying fruit fly species that face the challenges presented in this application. This study contributes to entomologists interested in automating the classification of this important group. In a broader context, this contribution is part of the broader picture of automating biodiversity monitoring, which, in some cases, requires specialised algorithms to provide an accurate representation of biodiversity and therefore facilitate the implementation of Caughley's management actions. So, the presented exploration of potential approaches provides insights towards this goal.

We identified two primary limitations of the presented study. The first relates to the challenge itself, where algorithm generalisation is affected by data scarcity, as evidenced by the high variance in the performance metrics. Thus, further work to obtain additional images of this rare species could improve the performance of learning algorithms and generalisation. The second limitation concerns the fact that augmentation algorithms do not necessarily reflect real biological variation, but risk overfitting to the artefacts of the limited original dataset. Finally, we highlight that the overall accuracy is only appropriate as a performance metric in balanced datasets, so future work should include precision/recall, F1, or per-class confusion matrices.

2.3.6 Conclusion

In this this chapter, we proposed a new approach to classify species of the *Anastrepha pseudoparallela* group using classical computer science techniques with machine learning and computer vision methods. The main challenges faced by our work relate to the group's rare species characteristics, which results in imbalanced scarce data. We showed that combining the proposed features collected from the wing's node structure and the RGB information from specific polygons formed by specialists can better classify those species than the traditional VGG16-PCA architecture based on the various learning algorithms tested. Moreover, we showed that the use of coloured images with SMOTE and Random Forests algorithms provided a higher classification performance based on these species. This study serves as an initial exploration of the classification of rare *Anastrepha* species and can serve as a basis for classifying fruit fly species of other groups that present scarce data availability and class imbalance.

CHAPTER 3

Statistical machine learning applied to animal control

The main contributions of this chapter to Caughley's management action to decrease excessive populations are based on the design of a new non-black-box machine learning method to detect animal outbreaks and analyse the effect of animal dynamics on the outbreaks. Moreover, this Chapter illustrates the use of statistical machine learning techniques as a promising method for Integrated Pest Management (IPM) by focusing on the examples of the proposed learning algorithm for insect pests. Motivated by this problem, a new approach was introduced to predict insect abundance, considering the time series dependencies among insect and climate time series. Finally, this Chapter focused on researchers interested in utilising Statistical Machine Learning methods to improve their decision-making in IPM practices.

Entomology is the central focus of this Chapter, where insect populations and concerns about reducing economic damage caused by Aphid species illustrate the new non-black-box machine learning method and the new approach presented in the second topic. We start with the introduction of the Pattern-Based Prediction (PBP) method. Then, we introduce a new approach for estimating animal abundance based on climate covariates, machine learning, and time series embedding.

3.1 Pattern-Based Prediction of Population Outbreaks

As insect outbreaks provide a clear illustration of the target Caughley's management action, this application will continue to focus on the entomological context to propose a new machine learning method to predict insect outbreaks. The complexity and practical importance of insect outbreaks have made predicting them a focus of recent research. We propose the Pattern-Based Prediction (PBP) method for predicting population outbreaks. This method treats outbreaks as a binary problem, where the classes are the absence or presence of the studied event. The proposed method uses information from previous time series values preceding an outbreak as predictors of future outbreaks, which can be helpful for monitoring pest species. We illustrate the methodology using simulated datasets and an aphid time series obtained in wheat crops in Southern Brazil. We obtained an average test accuracy of 84.6% in simulation studies using stochastic models and 95.0% for predicting outbreaks using a time series of aphid counts in wheat crops in Southern Brazil. Our results demonstrate the feasibility of the PBP method for predicting population outbreaks. We benchmarked our results against established state-of-the-art machine learning methods: Support Vector Machines, Deep Neural Networks, Long Short-Term Memory, and Random Forests. The PBP method yielded competitive performance, with higher true-positive rates in most comparisons, while providing interpretability rather than being a black-box method. It is an improvement over current state-of-the-art machine learning tools, especially for non-specialists, such as ecologists, who aim to use a quantitative approach to pest monitoring.

3.1.1 Introduction

Automated systems for syndromic surveillance have been reported in many studies in different contexts, such as public health aiming to predict disease outbreaks and also agricultural pests [Madden and Wheelis, 2003, Buckeridge, 2007, Büntgen et al., 2020, Bright et al., 2020, Burkom et al., 2021]. These studies have demon-

strated many possibilities for predicting outbreaks based on population dynamics, sampling methods, outbreak frequency, and threshold analysis. Their results have demonstrated potential to help public health actions, based on the interpretation of results provided by these algorithms, using different data sources containing simulated and observed outbreaks [Buckeridge, 2007, Chan et al., 2021].

Historically, the algorithms used to predict outbreaks involved classical time series methods, such as ARIMA-type models, seasonal models, and partial differential equations, among others [Buckeridge, 2007]. These tools positively impacted public health actions by enhancing the possibility of predicting disease outbreaks, but the applications of these methods are not restricted to this area. In quantitative ecology, these applications were expanded so that many authors started representing ecological phenomena with mathematical, statistical and machine learning methods [Sarah P. Otto, 2007, Odum, 2005, Ross, 1998]. One recent example is the application of supervised machine learning methods for predicting infestations of pine trees by a mountain pine beetle [Ramazi et al., 2021].

Examples of these applications are the representation of biological systems and the interactions between the species, such as predator-prey, host-parasitoid and competition models [Odum, 2005, Badkundri et al., 2019]. Among the taxonomic groups used to study these applications, insects expand the possibility of developing these methods with various models inspired by different problems. Examples are the LPA (larva-pupa-adult), host-parasitoid and other models related to herbivory [Sarah P. Otto, 2007]. These models are commonly used to study biological phenomena such as outbreaks. In addition, researchers can implement further statistical and mathematical modelling studies with insect time series data.

Insects depend on resource availability, as demonstrated by many studies involving time series [Nair, 2001, 2007, Santos et al., 2017, Lantschner et al., 2019]. The co-evolution between pests and plants reinforces this influence. It shows the complexity of this system, including biochemical strategies to avoid herbivory and the genetic plasticity of pests, enhancing their capability to obtain the necessary resources from plants [Wallner, 1987, Nair, 2001, 2007]. Pest dynamics are also based on physical and biological conditions, such as temperature, humidity, pre-

cipitation or irrigation, which strongly influence pest density [Wallner, 1987, Nair, 2001, Odum, 2005]. Biological factors such as mating system, life cycle, number of offspring per generation, mortality, and resource availability are also capable of influencing insect populations [Wallner, 1987, Godfray and Godfray, 1994, Nair, 2001, Hall et al., 2017]. In monoculture scenarios, crop phenology is followed by the presence of pests reinforcing that resource availability has a strong effect on the population dynamics in agroecosystems [Nair, 2001, 2007, Santos et al., 2017].

Insect outbreaks have frequently been documented in pest populations [Santos et al., 2017, Lynch, 2009, 2018]. They are important biotic disturbances in forests and agroecosystems [Wallner, 1987, Nair, 2001, Lantschner et al., 2019], since they may cause economic and/or ecological damage. Studies have shown possible explanations for outbreaks, such as abundant resources in monocultures, absence of predators or parasitoids, genetic factors, and pheromones produced by pests [Hall et al., 2017, Tao et al., 2012]. Biotic disturbances can be intensified by climate change and human activities, which makes it essential to study how they influence pest outbreaks [Volney and Fleming, 2000, Sharma and Dhillon, 2018, Phophi et al., 2019]. One example of how human activities can affect outbreaks is the occurrence of bark beetles in temperate forests. These insects have devastated a large area of pine trees in the continental United States [Negron et al., 2008].

In Brazil, there are several examples of pest outbreaks in forests and crops, as for example *Eucalyptus* with *Thyrinteina arnobia* and *Stenalcidia sp* (Geometridae) [Zanuncio et al., 2006]; black wattle with *Oncideres impluviata* (Cerambycidae) showing annual outbreaks in the state of Rio Grande do Sul [Ono et al., 2014]; soybean with *Chrysodeixis includes* and *Anticarsia gemmatalis* (Lepidoptera: Noctuidae) also exhibiting high frequencies [Bueno et al., 2010, Santos et al., 2017]. Studies focused on the species mentioned above show that outbreaks occur suddenly because of different natural effects that can result in increased population densities [Nair, 2001, 2007, Santos et al., 2017, Lantschner et al., 2019]. However, the exact reason why an insect species population suddenly increases in number is still an open question [Ekholm et al., 2019]. Outbreak occurrence forecasting turns out to be an arduous task requiring a large amount of man-hours, extensive field work and different types of specialised equipment. Entomologists tradition-

ally have been using different interventions to reduce economic damage, mainly in agriculture. The most common method traditionally used for this task is to define an economic threshold level [Stern et al., 1959, Onstad, 1987].

Nowadays, the possibilities of actions against insect damage in crops can be found in the IPM domain, which briefly consists of employing biological, physical, chemical and genetic approaches to reduce the population densities of a pest [Stern et al., 1959, Goodell, 2009]. The economic threshold concept has been coupled with more complete analyses involving control functions given by the crop and pest population information [Mitchell et al., 2004, Dun et al., 2009, Tinsley et al., 2013]. When outbreaks are frequently recorded in insect populations, the probability that their population size is bigger than the economic-injury level increases, causing secondary outbreaks [Goodell, 2009]. Given that this biological disturbance can occur suddenly and pest monitoring can be delayed, this density can be bigger than the economic threshold in the subsequent monitoring sample.

Different approaches have been proposed to address problems of this nature. One example is the Alert Zone Procedure (AZP) [Hilker and Westerhoff, 2007], which consists of scanning observations preceding population outbreak events to obtain profiles associated with these outbreaks. This method can extract meaningful information about outbreaks because previous densities before this event allow for the comprehension of ecological patterns. This method can be used as a basis to improve pest outbreak forecasting. However, it must be improved to deal with real-world problems. The effectiveness of these approaches remains an open question that is the focus of this research.

This chapter proposes the Pattern-Based Prediction (PBP) method to predict animal outbreak occurrence, which is an extension of the AZP based on statistical machine learning. For the purposes of this chapter, we consider an outbreak to have occurred when the population exceeds a given threshold level. We are interested in studying the pattern preceding such an outbreak; by pattern, we mean a sequence of population values in the times leading up to the outbreak occurrence. We will introduce more formal definitions in the following section. We begin by describing the method and then carry out simulation studies to assess the performance of

our method under different conditions. Finally, we illustrate our proposed method using a dataset obtained from a pest management system aimed at monitoring aphids, which are important pests for many different cultures – such as wheat, barley, and mustard [Kranti et al., 2021] – and discuss the feasibility of applying it to the context of pest management.

3.1.2 Methods

Generating patterns

Let x_t represent the population size of a particular species at time point t, $t = 1, \ldots, T$. Initially, we set a population size threshold x^* such that when $x_t \geq x^*$ we have a population outbreak at time t. We then implement the AZP, as proposed by Hilker and Westerhoff [2007]. This method consists of scanning observations to identify each outbreak event i, $i = 1, \ldots, I$, that occurred at time point t_i, based on the value of x^*, and collecting the m observations that precede them, forming a vector $\mathbf{p}_i^T = \{p_{i1}, p_{i2}, ..., p_{im}\} = \{x_{t_i-1}, x_{t_i-2}, ..., x_{t_i-m}\}$ per event. If $t_i - m < 1$, event i is ignored. After that, we group all population dynamics patterns that precede these events as the matrix

$$\mathbf{P} = \begin{bmatrix} x_{t_1-1} & x_{t_1-2} & \cdots & x_{t_1-m} \\ x_{t_2-1} & x_{t_2-2} & \cdots & x_{t_2-m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{t_I-1} & x_{t_I-2} & \cdots & x_{t_I-m} \end{bmatrix} = \begin{bmatrix} \mathbf{p}_1^T \\ \mathbf{p}_2^T \\ \vdots \\ \mathbf{p}_I^T \end{bmatrix}, \tag{3.1}$$

where I represents the total number of identified patterns. See Figure 3.1(a) for a plot of all rows of a hypothetical $\mathbf{P}$ matrix. Note that time series pre-processing may be carried out prior to obtaining the pattern matrix P. For instance, in Section 3.5 we compare the performance of our method using the raw time series and a pre-processed series using the Empirical Mode Decomposition method [Kim et al., 2012]. Overall, this method decomposes a time series into intrinsic mode functions (IMFs), and the key characteristic of this additive decomposition is the separation of deterministic components from stochastic elements [Huang et al., 1998].

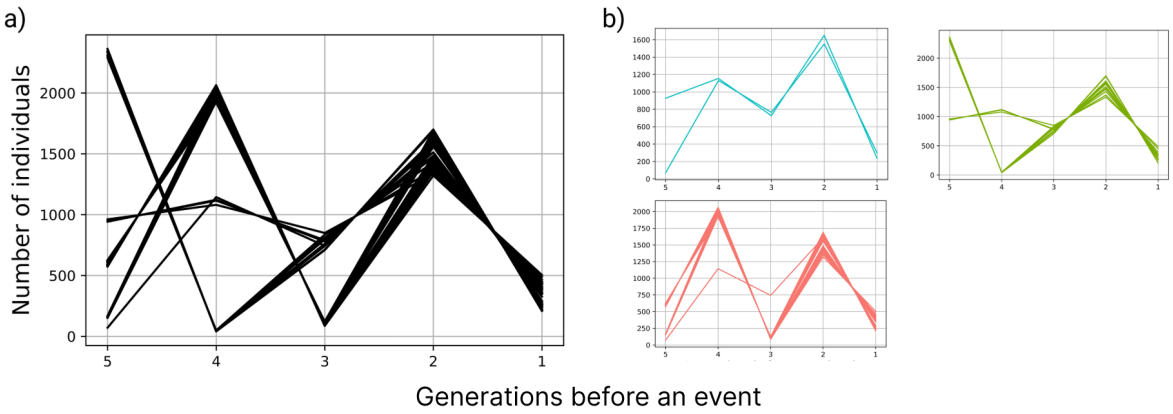

Figure 3.1: a) The representation of patterns $\mathbf{p}_i$ within the matrix $\mathbf{P}$ that precede an outbreak event, using $m = 5$. b) The respective cluster matrices $\mathbf{P}'_c$ obtained using $d^*_{\text{cluster}} = 0.4$. These patterns were obtained from time series data simulated from a Ricker map, with $r = 3$ and $K = 1000$, $x_1 = 200$ and 1000 observations. The population size threshold for the outbreak event was set as $x^* = 2224$ representing the 90% percentile of the data.

Then, if patterns i and j $(i \neq j)$ are sufficiently similar, we group them in the same cluster. We do this based on the association metric

$$d(\mathbf{p}_i, \mathbf{p}_j) = \frac{1}{c(\mathbf{p}_i, \mathbf{p}_j) + 1}, \tag{3.2}$$

where $c(\mathbf{p}_i, \mathbf{p}_j) = \sum_{k=1}^{m} \frac{|p_{ik} - p_{jk}|}{|p_{ik}| + |p_{jk}|} > 0$ is the Canberra distance [Androutsos et al., 1998, Ehsani and Drabløs, 2020] between two vectors, where $|\cdot|$ is the Euclidean norm. This distance is appropriate for non-negative count data [Androutsos et al., 1998]. Note that when $c(\mathbf{p}_i, \mathbf{p}_j) \to \infty$, then $d(\mathbf{p}_i, \mathbf{p}_j) \to 0$, and as $c(\mathbf{p}_i, \mathbf{p}_j) \to 0$, then $d(\mathbf{p}_i, \mathbf{p}_j) \to 1$.

To define the similarity of patterns we set the value d^*_{cluster}, representing the minimum association metric for considering $\mathbf{p}_i$ similar to $\mathbf{p}_j$. This yields the cluster matrices $\mathbf{P}'_c$, $c = 1, \ldots, C$, that include patterns which are similar to one another. To obtain these, we start with pattern $\mathbf{p}_1$, which represents the first row of the matrix $\mathbf{P}$. We remove $\mathbf{p}_1$ from $\mathbf{P}$ and add it as the first row of $\mathbf{P}'_1$. After that we compute the association metric between $\mathbf{p}_1$ and all subsequent rows of

$\mathbf{P}$. If $d(\mathbf{p}_1, \mathbf{p}_j) \geq d^*_{\text{cluster}}$, we add pattern $\mathbf{p}_j$ as the last row of the cluster matrix $\mathbf{P}'_1$ and delete it from $\mathbf{P}$. We repeat this process to obtain the cluster matrices $\mathbf{P}'_c, c = 1, \ldots, C \leq I$, until there are no more rows left in $\mathbf{P}$ (Algorithm 1 in the supplementary material). Note that the order of $\mathbf{P}$ is important for the clustering procedure. See Figure 3.1(b) for an example where the generated patterns were split into $C = 3$ cluster matrices. The value specified for d^*_{cluster} governs the number of clusters generated. As $d^*_{\text{cluster}} \to 1$, $C \to$ tends to the number of distinct pattern vectors, I, while as $d^*_{\text{cluster}} \to 0$, $C \to 1$ (i.e. all patterns belong to the same cluster).

After obtaining the C cluster matrices

$$\mathbf{P}'_c = \begin{bmatrix} x_{t_1-1} & x_{t_1-2} & \cdots & x_{t_1-m} \\ x_{t_2-1} & x_{t_2-2} & \cdots & x_{t_2-m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{t_{l'_c}-1} & x_{t_{l'_c}-2} & \cdots & x_{t_{l'_c}-m} \end{bmatrix}, \tag{3.3}$$

where l'_c is the number of rows of $\mathbf{P}'_c$, we compute the vectors of means $\overline{\mathbf{p}}'_c$, containing the mean of each column for cluster matrix $\mathbf{P}'_c$, to form the rows of the matrix

$$\mathbf{P}'_{\text{means}} = \begin{bmatrix} \frac{1}{l'_1}\sum_{c=1}^{l'_1} x_{t_c-1} & \frac{1}{l'_1}\sum_{c=1}^{l'_1} x_{t_c-2} & \cdots & \frac{1}{l'_1}\sum_{c=1}^{l'_1} x_{t_c-m} \\ \frac{1}{l'_2}\sum_{c=1}^{l'_2} x_{t_c-1} & \frac{1}{l'_2}\sum_{c=1}^{l'_2} x_{t_c-2} & \cdots & \frac{1}{l'_2}\sum_{c=1}^{l'_2} x_{t_c-m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{l'_C}\sum_{c=1}^{l'_C} x_{t_c-1} & \frac{1}{l'_C}\sum_{c=1}^{l'_C} x_{t_c-2} & \cdots & \frac{1}{l'_C}\sum_{c=1}^{l'_C} x_{t_c-m} \end{bmatrix} = \begin{bmatrix} \overline{\mathbf{p}}'^T_1 \\ \overline{\mathbf{p}}'^T_2 \\ \vdots \\ \overline{\mathbf{p}}'^T_C \end{bmatrix}. \tag{3.4}$$

The matrix $\mathbf{P}'_{\text{means}}$ contains the information of all cluster matrices $\mathbf{P}'_c$, and is used for the prediction of a future event. Given a new collection of observations $\mathbf{x}_{\text{new}}$, with length m, we compute the association metric between $\mathbf{x}_{\text{new}}$ and each row of $\mathbf{P}'_{\text{means}}$. If any computed association is greater or equal to d^*_{pred}, shown in Eq. 3.5, the threshold for prediction, we predict that an event will occur. Finally d^*_{pred} is

defined for each row of $\mathbf{P'}_{\text{means}}$ as a function of l'_c, the number of patterns that generated each vector of means:

$$d^*_{\text{pred}} = f(l'_c) = d^*_{\text{base}} + \frac{(1 - d^*_{\text{base}})}{(l'_c)^\alpha}, \tag{3.5}$$

where d^*_{base} is the baseline value of the association metric (the smallest it is allowed to be) and α is a constant that changes the shape of the function f (see Figure 3.2). When $l'_c \to \infty$, we have that $d^*_{\text{pred}} \to d^*_{\text{base}}$, and as $l'_c \to 1$, also $d^*_{\text{pred}} \to 1$. This means that for predicting that a new event will occur, we would need a larger association between $\mathbf{x}_{\text{new}}$ and a particular $\overline{\mathbf{p}}'_c$ that was obtained from a small number of patterns.

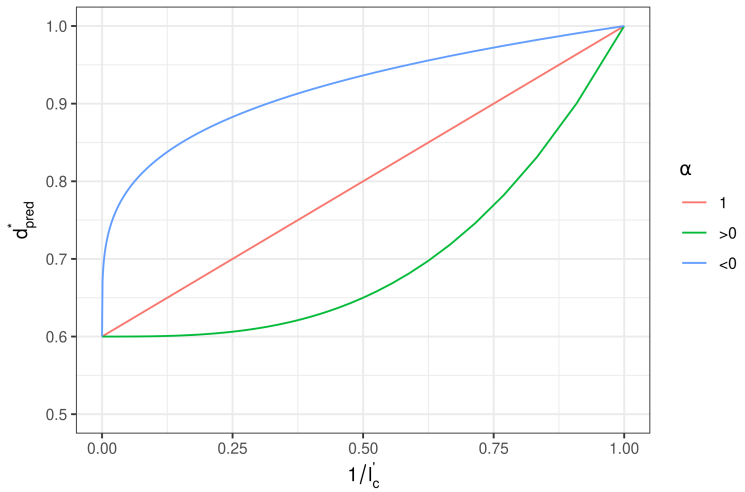

Figure 3.2: The threshold for prediction d^*_{pred}, calculated as a function of l'_c for $\alpha = 1$ (red curve), $\alpha = 3$ (green curve) and $\alpha = 0.25$ (blue curve), whilst fixing $d^*_{\text{base}} = 0.6$. The x-axis is represented as $1/l'_c$ to ease visualisation.

To summarise, the PBP method consists of the following steps:

1. Choose the value of the population size threshold x^*;
2. Set the values of m and d^*_{cluster};
3. Generate the pattern matrix $\mathbf{P}$;
4. Obtain the cluster matrices $\mathbf{P}'_c$ (algorithm presented in the supplementary material);

5. Compute the matrix $\mathbf{P}'_{\text{means}}$ from the column means of each cluster matrix $\mathbf{P}'_c$;

6. Set the values of d^*_{base} and α and obtain d^*_{pred} for each row of $\mathbf{P}'_{\text{means}}$;

7. Given a new collection of observations $\mathbf{x}_{\text{new}}$ compute the proposed association metric between $\mathbf{x}_{\text{new}}$ and each row of $\mathbf{P}'_{\text{means}}$;

8. If the computed association coefficient is greater than or equal to the value of d^*_{pred} associated with that row of $\mathbf{P}'_{\text{means}}$, predict that a new event will occur at the next time step; predict that it will not occur, otherwise.

A schematic diagram of this process is represented in Figure 3.3.

Choosing m, d^*_{cluster}, d^*_{base} and α via cross-validation

We propose the use of k-fold cross-validation to choose the values of m, d^*_{cluster} and α, such that the accuracy of the method is optimized. Here, the k-fold cross validation briefly consists of creating k groups of patterns $\mathbf{p}_i$ of the pattern matrix $\mathbf{P}$ and by removing the first group of patterns from $\mathbf{P}$, obtaining $\mathbf{P}'_{\text{means}}$ without using the information of this group, and carrying out the method to predict the occurrence of events based on the first group. After obtaining the method predictions based on this group, we compute

- true positives (TP): the number of times the method accurately predicted an event;
- true negatives (TN): the number of times the method accurately predicted there was no event;
- false positives (FP): the number of times the method incorrectly predicted an event;
- and false negatives (FN): the number of times the method incorrectly predicted there was no event.

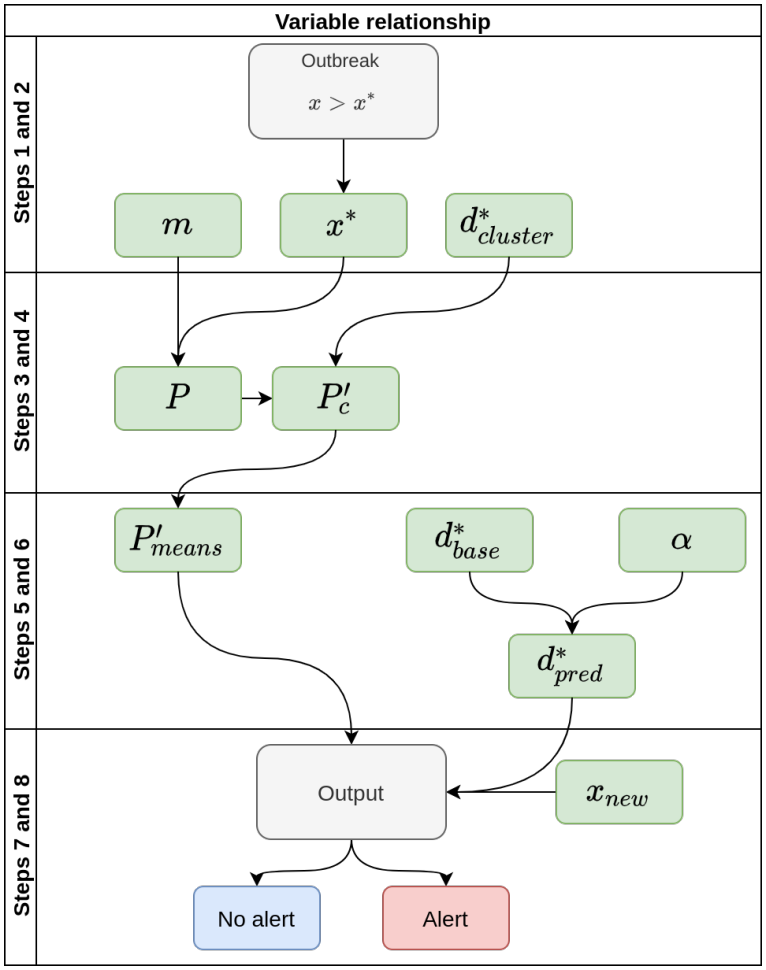

Figure 3.3: A schematic representation of the pattern-based method used to predict an outbreak based on time series data.

We use these values to obtain the accuracy, $\text{ACC} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}$, the true positive rate, $\text{TPR} = \frac{\text{TP}}{\text{TP}+\text{FN}}$, and the false positive rate, $\text{FPR} = \frac{\text{FP}}{\text{TN}+\text{FP}}$. We repeat this process for each group ending up with k values of these metrics. To measure the overall performance, we obtain the average of these metrics. Here, we carry out the analysis using $k = 5$.

To optimise the predictive power of the method, firstly we fix the values of m, d^*_{cluster} and α, and obtain different TPR and FPR values by varying d^*_{base}. The TPR and FPR can be plotted against each other to form a ROC curve [Hastie et al., 2004]. This curve is bounded between 0 and 1. For a method with good

predictive power, we expect the area under the curve (AUROC) to be close to 1. Let $g(m, d^*_{\text{cluster}}, \alpha)$ be an objective function that returns $-$AUROC based on the described method. Then, we use the Generalized Simulated Annealing method [Tsallis, 1988, Tsallis and Stariolo, 1996, Xiang et al., 1997, Xiang and Gong, 2000, Xiang et al., 2013, Mullen, 2014] to obtain the values of m, d^*_{cluster} and α that minimise g. This method speeds up the computation process compared to a grid search of the variables. Other methods may be used, such as Differential Evolution (see the supplementary materials).

To construct the ROC curve, we vary d^*_{base} from 0 to 1 using increments of 0.1 and calculate the AUROC using the trapezoid method [Liu and Pierce, 1994]. Finally, we apply one of two forms to choose d^*_{base}: the first is based on selecting a minimum threshold for the true positive rate (e.g. 0.8 or 0.9) and the second on selecting a maximum threshold for the false positive rate (e.g. 0.1 or 0.2). In summary, the method consists of the following steps:

1. Fix the values of m, d^*_{cluster} and α;
2. For different values of d^*_{base}, carry out k-fold cross validation and obtain the TPR and FPR for each fold, and compute the AUROC;
3. Choose m, d^*_{cluster} and α such that the AUROC is the largest;
4. Choose d^*_{base} based on the minimum TPR or maximum FPR that would be allowed in the study, based on the ROC curve with the largest area. (Note that the minimum TPR or maximum FPR allowed depends highly on the ecological system and objectives of the monitoring programme.)

Sensitivity analysis

We carried out a sensitivity analysis using simulated deterministic population dynamics, obtained from the Ricker map [Sarah P. Otto, 2007]:

$$x_{t+1} = x_t \exp\left[r\left(1 - \frac{x_t}{K}\right)\right], \tag{3.6}$$

where x_t denotes the population size of an organism at time t and the parameters $r > 0$ and $K > 0$ describe the intrinsic growth rate and carrying capacity of the

environment, respectively. We simulated 100 generations using $r = 3$, $K = 1000$ and an initial value of $x_1 = 200$.

In order to study the influence of different values of m and d^*_{cluster} on C (the total number of cluster matrices $\mathbf{P}'$) and the overall accuracy of the method, we used the simulated observations from the Ricker map, setting $x^* = 2224$ as the threshold for an outbreak event. This value corresponds to the 90% percentile of the simulated values from the deterministic Ricker map setting the parameter values as described above. We then employed the methodology described in the previous sections to obtain the matrices $\mathbf{P}'_{\text{means}}$ for m varying from 2 to 15 in increments of 1, and $d^*_{cluster}$ varying from 0 to 1 with increments of 0.1.

Method validation under stochastic conditions

To study the accuracy of the proposed method in predicting outbreak and extinction risk events under stochastic conditions, we simulated from three different approaches. The first included an additive Gaussian error $\varepsilon_t \sim \text{Normal}(0, \sigma^2)$ in the Ricker map, yielding the recurrence equation

$$x_{t+1} = x_t \exp\left[r\left(1 - \frac{x_t}{K}\right)\right] + \varepsilon_{t+1}. \tag{3.7}$$

Whenever the addition of the random noise term yielded $x_{t+1} < 0$, a new random noise value would be drawn from the normal distribution until $x_{t+1} > 0$, to ensure positive population sizes.

The second approach utilized a state-space formulation using a Poisson distribution with the mean term μ given by the Ricker recurrence equation, i.e.

$$X_1 \sim \text{Poisson}(\mu_1 = x_1) \tag{3.8}$$

$$X_{t+1}|X_t \sim \text{Poisson}\left(\mu_{t+1} = X_t \exp\left[r\left(1 - \frac{X_t}{K}\right)\right]\right), \tag{3.9}$$

from which all x_t values were drawn recursively. Finally, the third approach utilizes a state-space formulation based on a negative binomial distribution to accommodate overdispersion in the simulation study, i.e. $X_{t+1}|X_t \sim \text{Negative Binomial}(\mu_{t+1}, \phi)$. We estimated r and K based on real time series data of aphid counts in Southern Brazil for each model formulation, as well as the dispersion parameters σ^2 for the Gaussian model and ϕ for the negative binomial model.

Using the parameter estimates in Table 3.1, we simulated 20 samples of size 400 for each model. We also simulated 20 samples of size 400 using the negative binomial model with $\phi = 3$, to introduce a scenario with stronger overdispersion. We computed the accuracy, TPR and FPR by training the methods with the initial 80% observations and testing with 20% of the time series. Moreover, based on the ROC curve with the largest AUROC, we chose d^*_{base} using four methods:

1. 'TPR_08': choose the d^*_{base} value associated with the smallest TPR value that is equal to or greater than 0.8;
2. 'TPR_09': choose the d^*_{base} value associated with the smallest TPR value that is equal to or greater than 0.9;
3. 'FPR_01': choose the d^*_{base} value associated with the largest FPR value that is equal to or less than 0.1;
4. 'FPR_02': choose the d^*_{base} value associated with the largest FPR value that is equal to or less than 0.2.

Parameter	Model		
	Gaussian	Poisson	Negbin
r	0.15	0.28	0.57
K	224	310	370
σ^2	21,866	–	–
ϕ	–	–	1.2
AIC	5,226	34,913	4,296

Table 3.1: Parameter estimates obtained when fitting the Ricker state-space model to the aphid data assuming different distributions for the observation process, namely Gaussian, Poisson and negative binomial, as well as the Akaike Information Criterion (AIC) for each model fit. Negbin = negative binomial.

Analysis of case-study: Aphid data

To illustrate the predictive performance of our method, we use data obtained from an aphid monitoring programme implemented in Southern Brazil (State of Rio Grande do Sul, RS). These insects are considered as important pest species of

many crops. For instance, among the aphid species monitored by this programme, the species *Rhopalosiphum padi* and *Rhopalosiphum rufiabdominalis* are widely considered important pest species associated with winter cereals, and are found in the Eurasian region with a cosmopolitan distribution [Macfadyen and Kriticos, 2012]. Sampling was carried out weekly in an area of $5500m^2$ in a wheat culture region (Coxilha, RS, 710 m altitude, 28°11′42.8″ S and 52°19′30.6″ W), from 2011 to 2019, totalling 424 observations. The temperature and relative humidity data were monitored at the Passo Fundo weather station (28°15′ S, 52°24′ W, 684 m), located 10 km from the experimental area. The field was cultivated under a no-till system.

The species of aphids were monitored using Moericke traps (yellow tray, 45 cm long x 30 cm wide x 4.5 cm high), filled with a solution (2 L) consisting of water, 40% formalin (0, 3%) and detergent (0.2%). Each tray had three lateral holes (5 mm in diameter) close to the edge, protected by a thin screen to prevent leaks and loss of solid content during rain. Four traps were distributed at the borders of the crop rotation tests. The traps were levelled at approximately 20 cm from the floor with bricks. The crop rotation area was cultivated with cereals (oat, wheat, and triticale), radish and fallow during the Winter and in the Summer with soybeans, corn, and *Brachiaria* sp. Every seven days, the solid content of the trays was separated from the solution through the sieve and collected. The biological material was preserved in a glass bottle with 70% alcohol. Aphids and parasitoids were separated, identified, and counted under a stereomicroscope in the laboratory.

Monitoring is one of the bases for IPM. For aphids, their importance stands out mainly due to the ability of these insects to transmit viruses (Barley Yellow Dwarf Virus) to economically important crops, such as wheat, oats, barley and rye. The criterion to find a threshold (x^*) was the total number of aphids observed in the four traps. Usually, 10% of plants infested by aphids results in an economic threshold [Bell et al., 2015]. This percentage corresponds to 50 insects per trap, totalling 200 aphids, which is the threshold value used here to define an outbreak.

We selected 40%, 50%, 60% and 70% of the initial observations of the time series

for training and the complement as test sets to obtain the accuracy, true-positive rates and false-positive rates. We compared the performance of the PBP method with Random Forests (RF) [Breiman, 2001, Hastie et al., 2009] and Support Vector Machines (SVM) [Cortes and Vapnik, 1995, Crammer and Singer, 2001] algorithms. To obtain the algorithm performance using these competing methodologies, we created the matrix $\mathbf{P}$ with $m = 4$. The classification threshold of both algorithms were selected based on the same four criteria based on true and false positive rates used by the proposed method. Another matrix that did not contain outbreaks was generated with the same m observations before a threshold lower than the population size $x^* = 200$ that defines an outbreak of aphids in the study area.

For RF, we used 2 splitting predictors per tree (for $m = 4$), and a total of 1,000 trees. For SVM, we used the linear kernel. The implementation of the classification methods was carried out in Python using the library Scikit-learn [Van Rossum and Drake Jr, 1995]. Finally, we obtained the accuracy, true positive rate and false positive rate using the aforementioned training-test splits of the time series.

The PBP method is implemented through the `pypbp` package, which may be directly downloaded from the Python Package Index server or accessed through `https://pypbp-documentation.readthedocs.io`.

3.1.3 Results

3.1.3.1 Sensitivity analysis

We found that $d^*_{cluster}$ is proportional to the number of cluster matrices (C) created by the proposed method (see Figure 3.4(a)). However, as $d^*_{cluster}$ reaches values higher than 0.45, the parameter m did not influence the accuracy of our methods. It indicates that fixed values of m could be used when we use such values of $d^*_{cluster}$ (Figure 3.4(b)). These findings highlight the importance of using optimisation procedures to choose the appropriate value of $d^*_{cluster}$ for each study.

3.1.3.2 Method validation under stochastic conditions

The accuracy of our method for predicting population outbreaks obtained from the stochastic simulation scenarios using the raw time series data and pre-processing

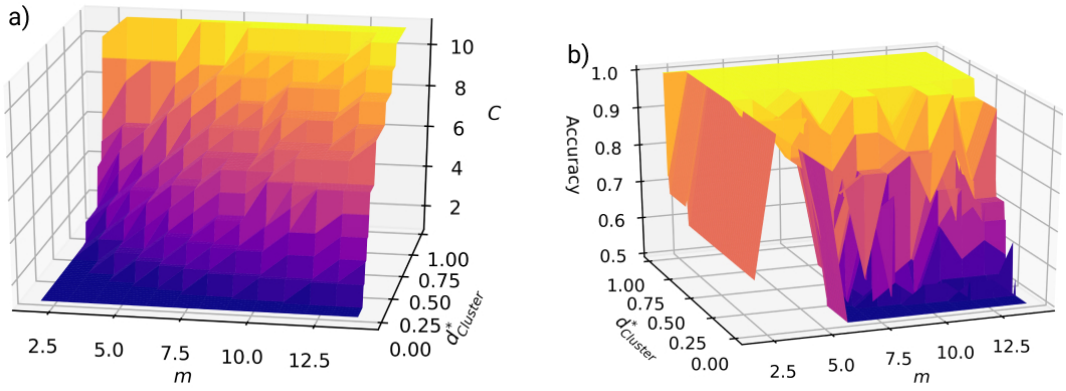

Figure 3.4: The effect of m and $d^*_{cluster}$ on (a) C (i.e. number of cluster-matrices $\mathbf{P}'_c$), and on (b) the accuracy of the proposed method. These results were obtained from time series data simulated from a Ricker map, with $r = 3$ and $K = 1000$, with $x_1 = 200$. The population size threshold for the outbreak event was set as $x^* = 2224$ representing the 90% percentile of the data.

the data using Empirical Mode Decomposition (EMD) considering all models were, respectively, on average 73.8% with a standard deviation of 23.4% and 73.2% with a standard deviation of 24.1% (see Figure 3.5). The average FPR obtained was 25.3% with a standard deviation of 29.0% and 25.1% with a standard deviation of 28.7%. Finally, the average TPR were 53.2% with a standard deviation of 40.9% and 55.6% with a standard deviation of 41.9%. Therefore, we found that there are no differences in performance when pre-processing the data using EMD.

Also, considering the influence of the model, we found that the rank of models in which our method produced higher performance is, respectively, starting with the best one, the negative binomial ($\phi = 1.2$), Gaussian, negative binomial ($\phi = 3$) and Poisson stochastic models. On average, we obtain an accuracy of 84.6% with a standard deviation of 20.5%, a false positive rate of 14.9% with a standard deviation of 24.2% and a true positive rate of 59.6% with a standard deviation of 42.0% for the negative binomial ($\phi = 1.2$) model. Considering the Gaussian model, we obtained an accuracy of 75.2% with a standard deviation of 18.0%, a false positive rate of 22.9% with a standard deviation of 22.2% and a true positive rate of 55.5% with a standard deviation of 37.3%. This finding makes our method a promising prediction tool since we got good results even when using stochastic

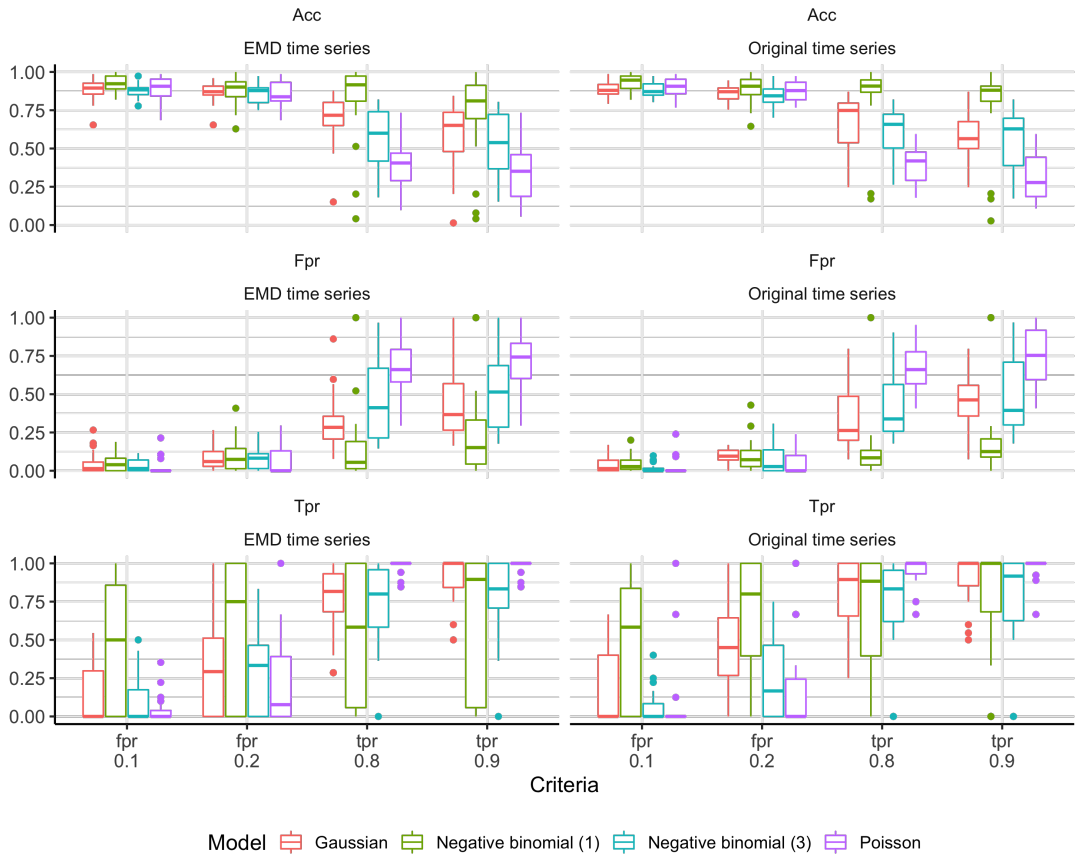

Figure 3.5: Accuracy simulation results, TPR (True Positive Rate) and FPR (False Positive Rate) using the raw simulated time series and pre-processed series using Empirical Mode Decomposition (EMD). In both scenarios four methods were used to choose d^*_{base}: based on a maximum FPR (0.1 and 0.2) or a minimum TPR (0.8 and 0.9).

approaches to simulate the data.

3.1.3.3 Analysis of case-study: Aphid data

To predict the threshold representing an outbreak for the aphid population dynamics (Figure 3.6), we select $x^* = 200$ considering the number of species collected in the four traps of the monitoring system, which was related to the 10% of plants infected by aphids resulting in the economic threshold. Applying PBP using different training data obtained from percentages of the initial observations of the

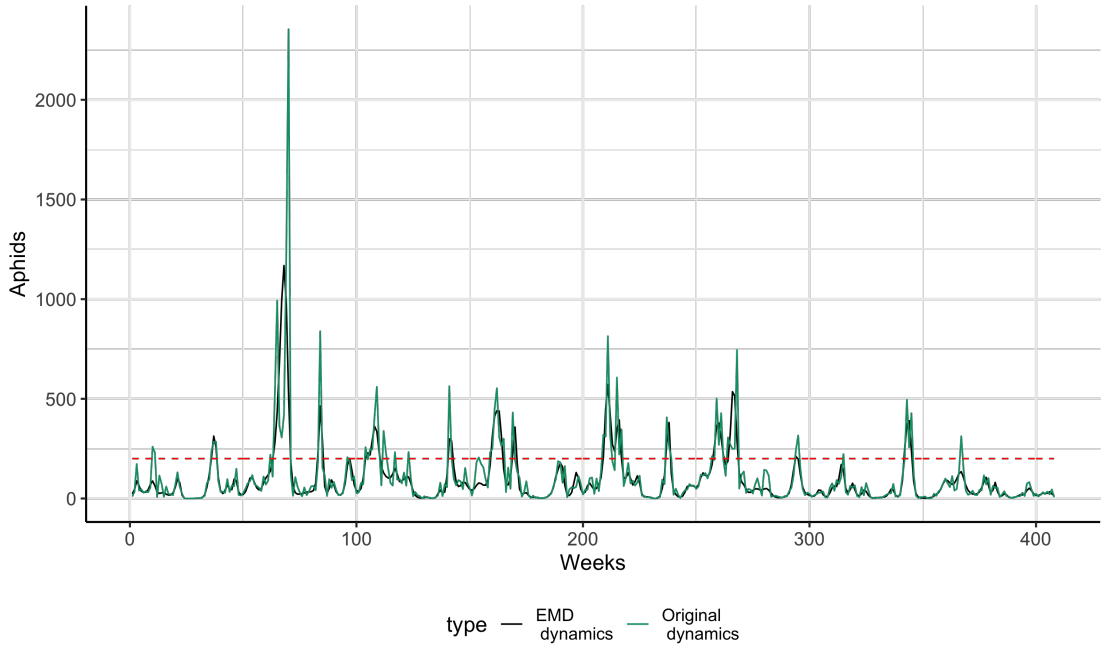

Figure 3.6: The time series represents the total aphids collected within the four traps on the monitoring system on time. The red line represents the threshold $x^* = 200$, the green line is the original time series, and the black line is the result of the empirical mode decomposition method.

time series of aphids, the accuracy values were higher than 70% regardless of the percentage of training using the original aphid time series.

Table 3.2: Prediction accuracy, true-positive rate (TPR) and false positive rate (FPR) obtained from the Pattern-Based Prediction (PBP) and competing methods Random Forests (RF) and Support Vector Machines (SVM) with $m = 4$ (i.e. 4 observations before the event). Classification thresholds were selected based on four criteria: $FPR \leq 0.1$, $FPR \leq 0.2$, $TPR \geq 0.8$, and $TPR \geq 0.9$. All methods were carried out using training sets with $40\%, 50\%, 60\%, 70\%$ and 80% of the initial observation of the aphid time series.

Metrics	Train percentage	PBP ($FPR \leq 0.1$)	PBP ($FPR \leq 0.2$)	PBP ($TPR \geq 0.8$)	PBP ($TPR \geq 0.9$)	RF ($FPR \leq 0.1$)	RF ($FPR \leq 0.2$)	RF ($TPR \geq 0.8$)	RF ($TPR \geq 0.9$)	SVM ($FPR \leq 0.1$)	SVM ($FPR \leq 0.2$)	SVM ($TPR \geq 0.8$)	SVM ($TPR \geq 0.9$)
Accuracy	0.4	0.81	0.81	0.71	0.49	0.88	0.80	0.80	0.73	0.89	0.89	0.68	0.68
	0.5	0.87	0.87	0.87	0.87	0.88	0.81	0.85	0.70	0.89	0.89	0.32	0.32
	0.6	0.90	0.88	0.83	0.83	0.88	0.81	0.82	0.72	0.91	0.83	0.89	0.83
	0.7	0.94	0.94	0.92	0.87	0.89	0.81	0.84	0.78	0.93	0.90	0.90	0.90
	0.8	0.95	0.90	0.90	0.90	0.90	0.82	0.84	0.84	0.96	0.82	0.96	0.96
TPR	0.4	0.52	0.52	0.78	0.96	0.70	0.78	0.81	0.93	0.67	0.67	0.96	0.96
	0.5	0.18	0.18	0.18	0.18	0.72	0.84	0.80	0.92	0.72	0.72	1.00	1.00
	0.6	0.12	0.69	0.87	0.87	0.62	0.81	0.81	0.94	0.69	1.00	0.88	1.00
	0.7	0.28	0.28	0.43	0.71	0.71	0.86	0.86	1.00	0.71	1.00	1.00	1.00
	0.8	0.50	0.75	0.75	0.75	0.75	1.00	1.00	1.00	1.00	1.00	1.00	1.00
FPR	0.4	0.15	0.15	0.30	0.57	0.10	0.19	0.21	0.30	0.08	0.08	0.36	0.36
	0.5	0.05	0.05	0.05	0.05	0.09	0.20	0.14	0.33	0.08	0.08	0.78	0.78
	0.6	0.01	0.09	0.18	0.18	0.09	0.19	0.17	0.30	0.06	0.19	0.11	0.19
	0.7	0.02	0.02	0.05	0.11	0.10	0.19	0.16	0.23	0.05	0.11	0.11	0.11
	0.8	0.03	0.09	0.09	0.09	0.10	0.19	0.16	0.16	0.04	0.19	0.04	0.04

Table 3.2 shows the performance of the PBP method compared to state-of-the-art machine learning methods, such as the commonly used Random Forest (RF) algorithm. The criterion of a false positive rate of at most 0.2 provides an accuracy of 90.0%, a false positive rate of 9% and a true positive rate of 75%. Moreover, our method using the criteria based on the true positive rate values of a minimum of 0.8 and 0.9 could obtain higher values of 96%. On the other hand, the method got an accuracy of 71% and a false positive rate of 32% in both cases. In practice, a false positive would result in using pest control techniques unnecessarily, while a false negative could result in failing to control the pest, which may cause economic damage due to an outbreak occurring. Considering the criterion of a false positive rate of at most 0.2, the SVM algorithm, achieved an accuracy, true positive rate, and false positive rate of 96%, 100%, and 4%, respectively. Finally, the RF achieved an accuracy, true positive rate, and false positive rate of 82%, 100%, and 19%, respectively. Our method presented a competitive performance with regard to false positive rate in this scenario. However, the other learning algorithms achieve a higher true positive rate. Although the performance of the proposed algorithm does not surpass classical machine learning algorithms, the proposed method offers greater interpretability compared to other learning algorithms.

3.1.4 Discussion

We proposed the Pattern-based Prediction (PBP) method and analysed its sensitivity and performance based on simulation studies and real data. We applied the method to a time series of aphids in wheat crops in Southern Brazil. The creation of cluster matrices and subsequent sensitivity analysis carried out here is inspired by and builds on the studies conducted by Hilker and Westerhoff [2007]. By using the AZP as a basis to create $\mathbf{P}'_c$, we enhance the information which can be extracted from the population dynamics of any species of interest. This process can extract the different population states using the dynamics obtained from monitoring programmes. Also, by grouping these states into different cluster matrices, we can observe the frequency of each pattern type occurring before the outbreak. The process of clustering will help us to perform the outbreak classification based on the different pattern types contained in $\mathbf{P}'_c$. These results reflect the accuracy of 100% obtained for some parameter regions in our sensitivity study. The re-

sults improved dramatically when we fully optimised our choice for the parameter values, even when subject to stochastic effects.

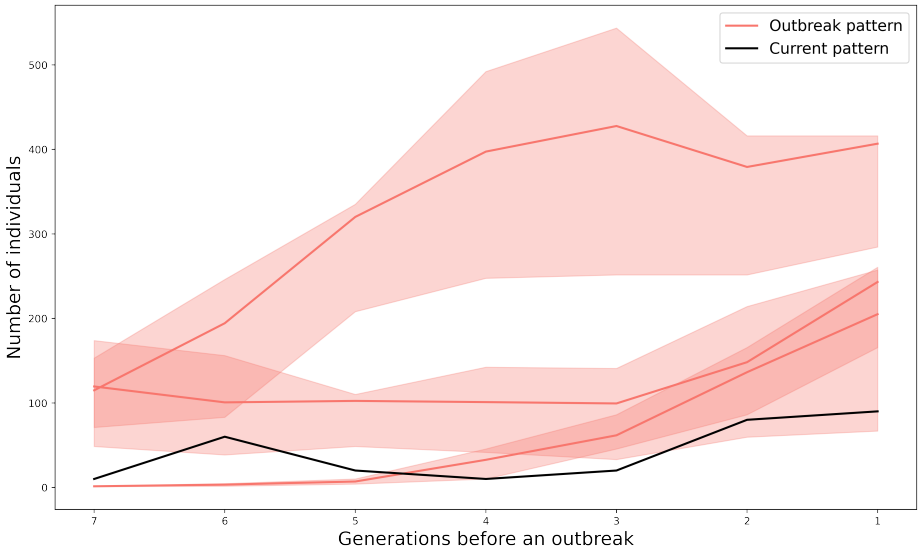

Figure 3.7: Outbreak patterns obtained from the proposed method using the aphid time series. Each of the red lines represents a row of the $\mathbf{P}'_{\text{means}}$ matrix. The intervals are the 25% and 75% percentiles of the patterns that generated each vector of means. The black line represents an observed series ($\mathbf{x}_{\text{new}}$), for which the association metric d is calculated between each of the three identified patterns. If $d > d^*_{\text{pred}}$, then the PBP method would classify $\mathbf{x}_{\text{new}}$ as preceding an outbreak event. The calculated d^*_{pred} values for the three patterns were 0.37, 0.61 and 0.42, whereas the association metrics between $\mathbf{x}_{\text{new}}$ and each pattern were 0.20, 0.26 and 0.15, respectively. Therefore, $\mathbf{x}_{\text{new}}$ would be classified as not preceding an outbreak.

With respect to interpretability, when using RF, it is possible to obtain variable importance. However in this case, this will tell us which previous steps were most important when predicting outbreaks, not necessarily how they relate to its occurrence. However, each hyperparameter in the PBP framework provides a clear interpretation, and we are able to create visual representations of the patterns that occurred before the outbreak (the $\mathbf{P}'_{\text{means}}$ matrix). For instance, Figure 3.7 displays the three patterns in $\mathbf{P}'_{\text{means}}$ obtained from employing PBP using the optimised

hyperparameter values for the aphid data using 50% of the time series for training.

In addition, the number of patterns encountered in each cluster matrix presents the importance of each clustered pattern for predicting animal outbreaks. The parameter α shows how relevant each group of patterns is for predicting an outbreak. Also, we can assess the importance of clustering the pattern matrix by looking at the estimate of d^*_{cluster}. Larger values typically indicate fewer recognised patterns in $\mathbf{P}'_{\text{means}}$. The m hyperparameter shows the number of previous observations required to provide a classification based on our method, so it provides a clear interpretation for ecologists and farmers in terms of how far in the past to watch for when identifying patterns. The d^*_{base} hyperparameter informs the minimum degree of similarity that is required to classify an outbreak, based on previously observed patterns. Therefore, not only is the PBP method competitive when compared to state-of-the-art machine learning methods, it is also interpretable, and brings descriptive advantages combined with its predictive power.

The introduction of the PBP method provides a venue for further exploration of the technique by including covariates for improving the occurrence of outbreaks. In future works, we will explore other statistical learning methods that can be used to improve the methodology in a scenario with climactic covariates. For example, methods such as convergent cross mapping, PCMCI, empirical dynamic modelling, or other machine learning models could be useful and help take nonlinear interactions into account Pianosi et al. [2016], Runge et al. [2019a]. Also, this study will serve as a basis for developing future animal monitoring programs enhanced with outbreak detection based on the PBP method.

3.2 Forecasting insect abundance using time series embedding and machine learning

In this application, the contributions to the studied management action involve a different perspective on insect outbreaks. Instead of classifying the events as in the previous application, we now present a new approach to predict insect abundance. The main challenge in this application is the open question of developing accurate interventions for insect control based on insect-monitoring systems. A possible

solution to enhance decision-making is to apply forecasting methods to predict insect abundance. However, another layer of complexity is added when other covariates are considered in the forecasting, such as climate time series collected along the monitoring system. Multiple combinations of climate time series and their lags can be used to build a forecasting method. Therefore, we propose a new approach to address this problem by combining statistics, machine learning, and time series embedding. We used two datasets containing weekly time series of aphids and climate data collected over eight years in two municipalities in Southern Brazil. We conduct a simulation study using a probabilistic autoregressive model with exogenous time series, based on Poisson and negative binomial distributions, to evaluate the performance of our approach. We pre-processed the data using our newly proposed approach and more straightforward approaches commonly used to train learning algorithms. We assess the performance of the selected algorithms by examining the Pearson correlation and Root Mean Squared Error obtained from one-step-ahead forecasting. Based on Random Forests, Lasso-regularised linear regression, and LightGBM regression algorithms, we showed the feasibility of our novel approach, which yields competitive forecasts while automatically selecting insect abundances, climate time series, and their lags to aid forecasting.

3.2.1 Introduction

Yield loss is an example of the impacts of the insect outbreak due to the consumption of plants by the pest species. Arthropod pests are responsible for 20% of global annual crop losses [Mateos Fernández et al., 2022]. Also, the transmission of diseases from insect-plant interaction [Perring et al., 1999, Smyrnioudis et al., 2001, Brown et al., 2002, Heck, 2018, Hoffmann et al., 2023] provides a clear example of how vital the correct management of insect outbreak is for avoiding economic damage.

When insect outbreaks start evolving in natural regions, they can disturb the food chain in the ecosystem, impacting biodiversity by reducing the population of other essential species [Müller et al., 2008, Lindroth et al., 2024]. As presented in session 3.1, there are several examples of insect pest species related to outbreaks in forests

and these examples motivate developing and implementing forecasting methods to prevent insect outbreaks. Especially methods that predict the best moment to proceed with interventions in the pest population. The possible types of interventions are well-studied in the IPM field. So, the vast number of alternative solutions, such as biological, physical, chemical and biotechnology control approaches [Mateos Fernández et al., 2022, Grijalva et al., 2024, Taggart et al., 2024] provides an excellent toolkit for growers; however, methods supporting accurate decisions for preventing insect outbreaks still lead to open research questions, helping the discovery of the appropriate solution to be applied in the field to reduce crop losses.

There are several approaches to implementing outbreak forecasting. One alternative is to predict the event using an algorithm to classify a binary problem (outbreak and non-outbreak) or provide forecasts for the abundance of insects. Based on event prediction, machine learning algorithms have demonstrated high performance for classifying insect outbreaks [Ramazi et al., 2021, Palma et al., 2023a, Jiang et al., 2024]. Also, machine learning has demonstrated promising results for forecasting insect abundance [Scavuzzo et al., 2018, Chen et al., 2019, Zhao et al., 2020, Rouabah et al., 2022, Ceia-Hasse et al., 2023, Kishi et al., 2023a]. The findings have shown that in some cases, machine learning methods achieve higher performance than traditional methods when analysing different types of time series [Khedmati et al., 2020, Spiliotis et al., 2020, Büttner and Rabe, 2021, Hamdoun et al., 2021, Maaliw et al., 2021, Masini et al., 2023]. These algorithms can include multiple exogenous time series to obtain forecasts of a target time series, and recent reviews highlighted the efficiency of high-dimension algorithms, such as Lasso-type, Random Forests and Ensemble-based algorithms [Masini et al., 2023]. These examples inspire the application of machine-learning algorithms to insect abundance forecasting in more depth.

Climate covariates are also collected over time in many insect monitoring systems. Since insects are poikilothermic organisms subject to meteorological variations on different temporal scales, climate coraviates may play an important role in their development. Meteorological variables such as temperature and rainfall are among the main abiotic factors that influence insect population dynamics. Among the complex pests affected by short- and long-term climate changes, aphids are the

most sensitive and are commonly used as a study model [Engel et al., 2022]. This brings additional opportunities to enhance insect abundance forecasting. However, it introduces more complexity to the analysis, considering the multiple combinations of lags from each climate time series [Brabec et al., 2014] that can be used as features of machine learning algorithms. To address this problem, this chapter introduces a novel approach for predicting insect abundance by combining statistics, machine learning, and time series analysis techniques. Our primary contribution is to provide a novel approach combining Takens' embedding theorem and Granger's causality to automatically select the lags of time series of insect abundances and climate variables to aid forecasting. We introduce a framework for understanding the causal effects of climate on insect abundance, considering multiple lags. Then, we use LightGBM, Lasso-regularised linear regression, and Random Forest algorithms to predict crop pest dynamics based on the forecast-focussed causal analysis. Finally, we combine these techniques for predicting insect abundance, illustrating our findings with two real-time series of aphid populations in the State of Rio Grande do Sul (RS) in Southern Brazil, and a simulation study based on Poisson and negative binomial autoregressive models with exogenous time series.

3.2.2 Methods

3.2.2.1 Reconstructing time series dependencies

Before proceeding with any learning strategy applied to temporal data, we must reconstruct it by unfolding time dependencies among observations [Mello and Ponti, 2018]. For instance, classical time series modelling approaches such as AR, ARMA, and ARIMA perform such reconstruction implicitly while modelling the influences that past observations have on current ones. In this context, we apply Takens' embedding theorem to explicitly reconstruct each observation $x(t)$ from a time series X with T observations, for all $t = 1, \ldots, T$, in a phase space coordinates $\mathbf{\Phi}$ in the form:

$$\phi_t = (x(t), x(t+\tau), \ldots, x(t+(\tau m - \tau))), \tag{3.10}$$

having m as the embedding dimension or the number of spatial axes, and τ as the time delay in between consecutive observations, finally ϕ_t corresponds to a position vector or state in a phase space $\mathbf{\Phi}$, i.e., $(\phi_1, \phi_2, \ldots, \phi_{T-(\tau m-\tau)}) \in \mathbf{\Phi}$.

In our particular scenario, given our interest in analysing cause-effect relationships among time series observations, we employed Granger's causality [Pearl et al., 2000, Shojaie and Fox, 2022] to map how a given exogenous or explanatory variable (another time series such as temperatures, rainfall, etc.) influences or anticipates events on the target time series (e.g. population size of insects) which is based on a set of past observations with the time delay $\tau = 1$, so that consecutive past observations are used to inform the machine learning approaches, following the suggestions of Rios and de Mello [2013]. The embedding dimension m can be estimated using Autoregressive models (AR) [Box et al., 2015]. Granger's causality allows the identification of the time delay between the exogenous series and the target series. In addition to such reconstruction, Granger's method requires the following additional steps:

1. Take the target time series Y (insect abundances over time) and employ first-order differences while it contains a relevant non-stationary component, which can be detected based on the Augmented Dickey-Fuller Test;

2. Take every exogenous time series X_i, where $i = 1, \ldots, I$ and I is the number of exogenous times series. For each X_i, ensure it is stationary by performing the same steps considered in the previous item;

3. Employ the cross-correlation function (CCF) on every pair on the stationary versions of an exogenous X_i and the target time series Y to measure the time delay for which the exogenous series has the greatest correlation (maximal correlation $MC(X_i, Y) = \underset{CCF(X_i,Y)}{\arg\max}$) with a future observation of the target time series;

4. Take time delay $\tau = 1$ and estimate the embedding dimension m by using the AR model on every time series (exogenous and target);

5. Employ Takens' embedding theorem to reconstruct all time series (exogenous and target), resulting in one data frame or panel per series;

6. The maximal correlation $MC(X_i, Y)$ is then used as a criterion to join data frames into a single dataset $\mathcal{D}$ and perform learning.

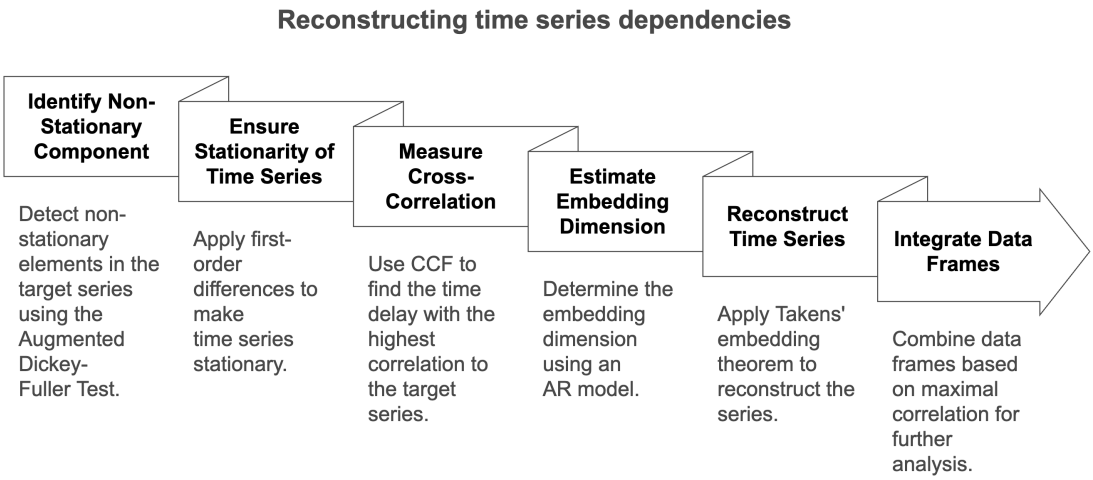

Figure 3.8: Flowchart of the necessary steps to reconstruct time series dependencies using Takens' embedding theorem and Granger's causality.

It is worth detailing how those data frames are merged based on the maximal correlation of every exogenous time series X_i with the target series Y described as $MC(X_i, Y)$. Suppose $MC(X_{\text{rainfall}}, Y) = -5$, this means every current value of Y, this is $Y(t)$, is most likely to depend on $X_{\text{rainfall}}(t-5)$. Consequently, they should be aligned before proceeding with modelling.

Suppose $m_{\text{rainfall}} = 3$ and $m_{\text{target}} = 2$, then we will have data frames built as follows: i) every row of the rainfall data frame will contain $(X_{\text{rainfall}}(t-2), X_{\text{rainfall}}(t-1), X_{\text{rainfall}}(t))$; and ii) every row of the target data frame will contain $(Y(t-1), Y(t))$. Now, taking into consideration $MC(X_{\text{rainfall}}, Y) = -5$, we must merge those data frames to produce $(X_{\text{rainfall}}(t-7), X_{\text{rainfall}}(t-6), X_{\text{rainfall}}(t-5), Y(t-1), Y(t))$ in which the three columns associated with the exogenous series were time displaced so that $X_{\text{rainfall}}(t-5)$ is used to bring as much information as possible to predict $Y(t)$. The same steps must be performed on all exogenous variables to obtain a single data frame $\mathcal{D}$, plugging all explanatory time series into our target series. Any machine learning method can be used to forecast $Y(t)$ after combining all explanatory time series into the target series based on the described method. Here, we present forecasting results based on three algorithms commonly used for this context: LightGBM, Random Forests and Lasso-regularised linear regression. Algorithm 2 illustrates the steps used by the proposed approach, and Figure 3.8

presents a flowchart highlighting the approach.

Algorithm 2 Time series reconstruction pseudo algorithm detailing the proposed approach.

Input: Target time series Y, exogenous time series $\{X_1, X_2, \ldots, X_i, \ldots, X_I\}$
Output: Reconstructed dataset $\mathcal{D}$ for learning
Step 1: Make Y stationary
    Apply first-order differences to Y until stationarity is achieved based on the Augmented Dickey-Fuller Test
for $i = 1$ to I **do**
    Step 2: Make X_i stationary
        Apply first-order differences to X_i until stationarity is achieved based on the Augmented Dickey-Fuller Test
end for
Step 3: Determine maximal correlation delays
for $i = 1$ to I **do**
        Compute Cross-Correlation Function (CCF) between X_i and Y
        Find time delay $\tau_i^* = \underset{CCF(X_i,Y)}{\arg\max}$
end for
Step 4: Estimate embedding dimension
    Set time delay $\tau = 1$
for each time series in $\{Y, X_1, \ldots, X_I\}$ **do**
        Fit an Auto-Regressive (AR) model to estimate embedding dimension m
end for
Step 5: Reconstruct time series using Takens' embedding
    Apply Takens' embedding theorem to each stationary time series with embedding dimension m and delay τ
    Create a data frame or panel for each reconstructed series
Step 6: Merge datasets based on maximal correlation
    Use $MC(X_i, Y)$ to join all reconstructed data frames into a single dataset $\mathcal{D}$
    Apply the chosen learning algorithm on the reconstructed dataset $\mathcal{D}$ to obtain forecasts.

3.2.2.2 Insect time series datasets

To compare the performance of the selected algorithms against each other, we used two real datasets of 211 observations each, including the total number of sampled aphids obtained from a monitoring system in Coxilha (S28°11′16.9″ W52°19′31.7″)

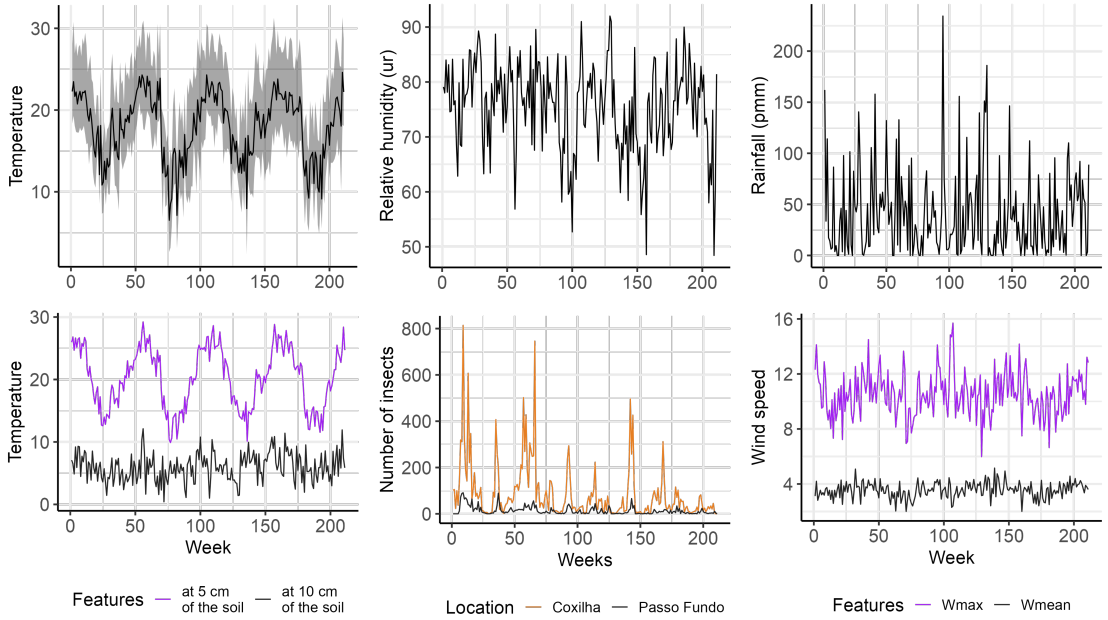

Figure 3.9: All exogenous and target time series used to illustrate the proposed approach.

and Passo Fundo (S28°13′36.6″ W52°24′13.4″) in the State of Rio Grande do Sul in Southern Brazil [Engel et al., 2022]. Each observation of the insect abundance time series contains climate covariates. In Table 3.3, we present a sample from the time series collected over the weeks related to the datasets and in Figure 3.9 we present a visualisation of the time series used in this chapter. Both regions use the same climate time series, given that the sampling areas are separated by approximately 12 kilometres. Thus, the main difference is the aphids' population dynamics.

3.2.2.3 Simulation study

The simulation study aims to gather insights into the requirements of the proposed method. Specifically, we aim to determine if the presence of climate time series impacts the performance of the proposed method when analysing the target time series. To conduct the simulation study, we begin by estimating the parameters of an Autoregressive model with exogenous time series (ARX) based on the Passo Fundo region dataset, including the insect abundance and climate time series. Let

Table 3.3: A sample of 4 weeks showing the features collected for Coxilha and Passo Fundo regions. The dataset contains the region (*region*), year (*year*), week (*w*), the temperature (*tmin* and *tmax*), rainfall (*pmm*), relative humidity (*ur*), wind speed (*wmax* and *wmean*), the temperature at 5 and 10 cm of the soil (*st5cm* and *st5cm*), and the aphid community total abundance (*aphids*).

region	*year*	*w*	*tmax*	*tmin*	*tmean*	*pmm*	*ur*	*wmax*	*wmean*	*st5cm*	*st10cm*	*aphids*
Coxilha	2015	1	28.03	18.35	22.27	162.00	79.10	12.31	3.12	7.09	25.96	102
Coxilha	2015	2	30.80	19.30	23.48	33.60	78.00	14.12	4.20	6.53	26.80	105
Coxilha	2015	3	26.94	18.03	21.67	114.10	84.00	11.97	2.81	4.87	24.69	23
Coxilha	2015	4	28.24	17.31	22.21	17.40	78.14	11.39	3.14	9.20	26.64	100
Passo Fundo	2015	1	28.03	18.35	22.27	162.00	79.10	12.31	3.12	7.09	25.96	0
Passo Fundo	2015	2	30.80	19.30	23.48	33.60	78.00	14.12	4.20	6.53	26.80	0
Passo Fundo	2015	3	26.94	18.03	21.67	114.10	84.00	11.97	2.81	4.87	24.69	0
Passo Fundo	2015	4	28.24	17.31	22.21	17.40	78.14	11.39	3.14	9.20	26.64	12

$y(t)$ be an observation from a discrete time series of insect densities Y, $x_i(t)$ be an observation from a time series of climate features X_i, where $i = \{1, \ldots, 9\}$ (i.e. the climate variables presented in Table 3.3). The autoregressive model of order p with exogenous time series can be written as:

$$\begin{aligned} y(t) = C + \omega_1 y(t-1) + \omega_2 y(t-2) + \ldots + \omega_p y(t-p) + \\ \theta_1 B^p x_1(t) + \theta_2 B^p x_2(t) + \ldots + \theta_9 B^p x_9(t) + \\ \theta_{10} x_1(t) + \theta_{11} x_2(t) + \ldots + \theta_{18} x_9(t) + \epsilon(t), \end{aligned} \tag{3.11}$$

where $\epsilon(t) \sim N(0, \sigma^2)$, $B^p x_i(t) = x_i(t-p)$ is the backshift operator, $\Omega = (\omega_1, \ldots, \omega_p)^\top$ are the autoregressive coefficients for the target time series Y, $\Theta = (\theta_1, \ldots, \theta_{18})^\top$ are the coefficients related to each climate time series X_i and their lags. To estimate the coefficients of the presented model, we first transformed the time series of insect densities by taking $Y' = \log(Y + 0.1)$. Then, all the coefficients were estimated using the package `fable` [O'Hara-Wild et al., 2023] for R software [R Core Team, 2025].

To explore different scenarios of insect dynamics, we used two distributions for generating Y and Y'. For the first scenario, we assumed that $Y(t+1) \sim \text{Poisson}(\lambda = \exp(y'(t) - 0.1))$. For the second distribution we assumed that $Y(t+1) \sim \text{Negative binomial}(\lambda = \exp(Y'(t) - 0.1), \theta = 1.8)$ to introduce overdispersion to the simulation. To investigate the impact of climate time series on the method performance, we first tested a scenario where we do not estimate any

of the Θ parameters, representing no influence of climate time series on the simulated insect abundance. For the second scenario, we estimated $\theta_1, \theta_2, \theta_3, \theta_4, \theta_5$, $\theta_{10}, \theta_{11}, \theta_{12}, \theta_{13}, \theta_{14}$ (related to *tmax*, *tmin*, *tmean*, *pmm*, *ur* climate time series and their lags p), representing the influence of half of the climate time series on the simulated insect abundance. Finally, for the third scenario, we estimated all Θ parameters related to the climate time series, representing the influence of all climate time series and their lags p on the simulated insect abundance. We used for each scenario $p = \{1, 3, 5\}$. For each combination of scenario and value of p, we generated 50 time series for both distributions, Poisson and negative binomial. Also, each simulated time series had 211 observations.

3.2.2.4 Forecasting performance of machine learning algorithms

To compare the performance of our novel method, we (i) used only the target time series (population of insects), with up to 3 or 6 lags behind to forecast future observations, (ii) used all climate time series as predictors with no lags, and (iii) all climate time series with target time series (population of insects) up to 3 or 6 lags behind to forecast future observations. We carried out an exploratory analysis using lags from 1 to 8 and decided to present the results utilising lags 3 and 6 to summarise the results without detracting from the central message of this chapter. We performed validation by obtaining one-step ahead forecasts for the entire time series (apart from the first 30 and 60 observations used for training the learning algorithms). For all time series of the case and simulation study, we trained Random Forests, Lasso-regularised linear regression and LightGBM algorithms for each approach (including all climate time series with no lags, time series reconstruction, taking the naive approach with up to 3 or 6-step lagged target series, and all climate times with up to 3 or 6-step lagged target series) and obtained their performance based on the Root Mean Squared Error (RMSE).

In addition, the proposed method provides a data frame $\mathcal{D}$ that selects climate, target time series and their lags during the reconstruction process, as described in Section 3.2.2.1. So, a data frame $\mathcal{D}$ with these features is created for every forecast during the one-step-ahead forecasting. To analyse the selection procedure that creates $\mathcal{D}$ on the forecasting performance of the proposed method, we collected

the number of selected features of $\mathcal{D}$ (the sum of the number of climate, target time series and their lags) for every forecast. Also, we collected the forecasting absolute error computed by the difference between the Random Forests' prediction and observed insect abundance. Finally, we analysed the correlation between the number of selected features of $\mathcal{D}$ (the sum of the number of climate, target time series and their lags) and the absolute error of the Random Forests algorithm.

We presented the forecasts of the Random Forests algorithm due to its inherent capability to incorporate non-linear relationships, allowing us to explore non-linear associations of most features provided by $\mathcal{D}$. Lasso-regularised linear regression would solely include linear associations, and the penalty provided by the L1 regularisation would decrease the number of features used in the forecasting, impacting the visualisation of the time series selection on the forecasting performance. The Light GBM could also be used based on these points. However, considering that the Random Forests algorithm has been commonly reported in papers with entomological applications [Chen et al., 2019, Valavi et al., 2021, Masini et al., 2023, Palma et al., 2023a], we solely report the results with this learning algorithm. Finally, we used the programming language R to implement all methods and the proposed approach. To allow for full reproducibility of the findings, we have made the code available at `https://github.com/GabrielRPalma/TimeSeriesReconstruction`.

3.2.3 Results

3.2.3.1 Case study

Figure 3.10 shows, as an overall result, that as the initial number of training samples increases, the RSME reduces in most scenarios. Considering all scenarios, approaches, and learning algorithms, the average RMSE reduced from 58.4 to 54.0 when the initial training sample increased. Overall, the average RMSE for Random Forests, Lasso-regularised linear regression and LightGBM were 54.3, 53.9 and 60.5. The average RMSE considering all datasets and algorithms for the proposed method, all climate time series with no lags, all climate time series with up to 3 or 6-lagged target series, and taking the naive approach with up to 3 or 6-lagged target series were, respectively, 50.1, 70.6, 54.3, 55.0, 52.2, and 55.0.

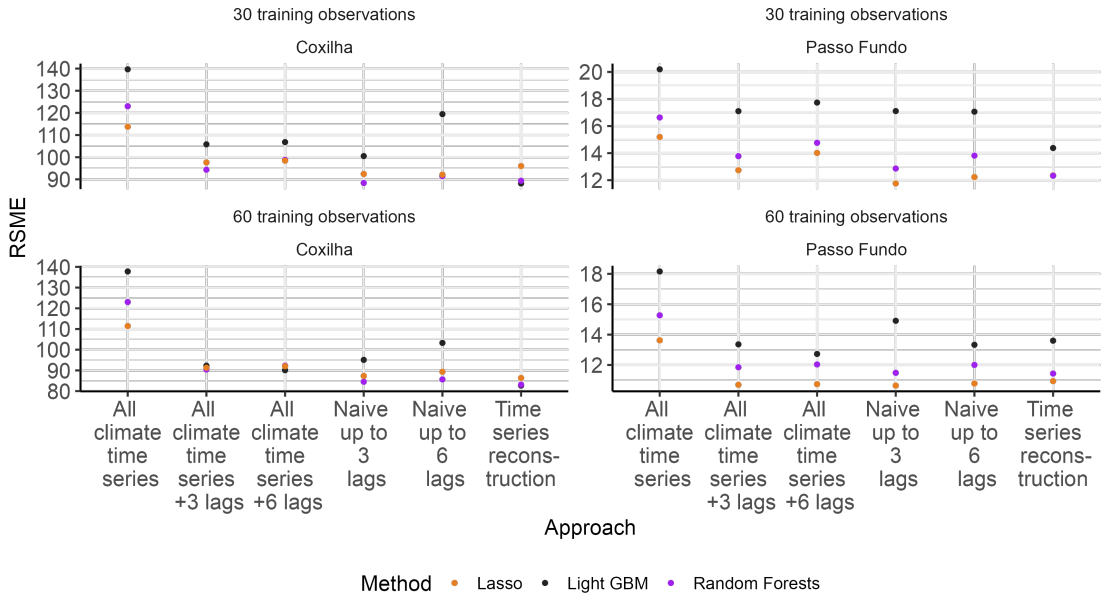

Figure 3.10: Root Mean Squared Error (RMSE) metric obtained by the Random Forests, Lasso-regularised linear regression, and LightGBM algorithms for each dataset (Coxilha and Passo Fundo with aphid's abundances), approach (including all climate time series with no lags, time series reconstruction, taking the naive approach with up to 3 or 6-step lagged target series, and all climate times with up to 3 or 6-step lagged target series) and the initial number of training samples used for each learning algorithm.

For Lasso-regularised linear regression, the average RMSE considering all datasets for the proposed method, all climate time series with no lags, all climate time series with up to 3 or 6-lagged target series, and taking the naive approach with up to 3 or 6-lagged target series were, respectively, 51.4, 63.5, 53.1, 53.8, 50.5 and 51.1. For Random Forests, the average RMSE considering all datasets for the proposed method, all climate time series with no lags, all climate time series with up to 3 or 6-lagged target series, and taking the naive approach with up to 3 or 6-lagged target series were, respectively, 49.1, 69.5, 52.6, 54.5, 49.3 and 50.7. The same scenarios for LightGBM resulted in 49.7, 79.0, 57.1, 56.8, 56.9, and 63.3.

Figure 3.11 shows that the Pearson correlation increases in most scenarios as the initial number of training samples increases. Considering all scenarios, approaches, and learning algorithms, the average correlation increased from 0.21 to 0.23 when

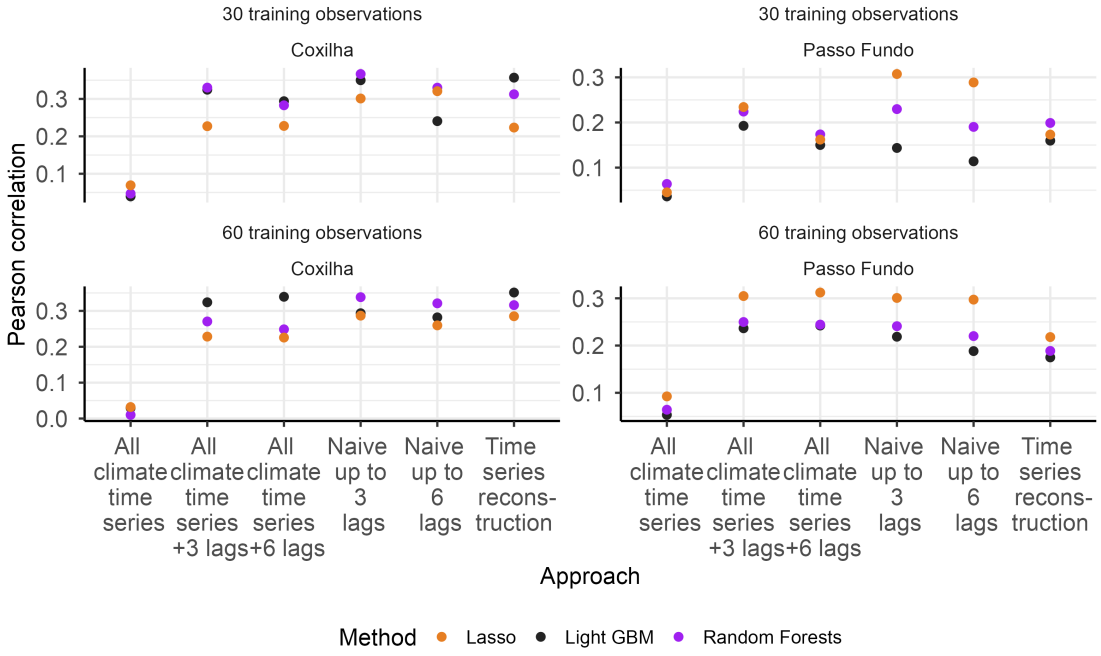

Figure 3.11: Pearson correlation metric obtained by the Random Forests, Lasso-regularised linear regression, and LightGBM algorithms for each dataset (Coxilha and Passo Fundo with aphid abundances), approach (including all climate time series with no lags, time series reconstruction, taking the naive approach with up to 3 or 6-step lagged target series, and all climate times with up to 3 or 6-step lagged target series) and the initial number of training samples used for each learning algorithm.

the initial training sample increased. Overall, the average correlation for Random Forests, Lasso-regularised linear regression and LightGBM were, respectively, 0.23, 0.23, and 0.21. The average correlation considering all datasets and algorithms for the proposed method, all climate time series with no lags, all climate time series with up to 3 or 6-lagged target series, and taking the naive approach with up to 3 or 6-lagged target series were, respectively, 0.25, 0.05, 0.26, 0.24, 0.28, and 0.25.

For Lasso-regularised linear regression, the average correlation considering all datasets for the proposed method, all climate time series with no lags, all climate time series with up to 3 or 6-lagged target series, and taking the naive approach with up to 3 or 6-lagged target series were, respectively, 0.22, 0.06, 0.25, 0.23, 0.30 and 0.29. For Random Forests, the average correlation considering all datasets for

the proposed method, all climate time series with no lags, all climate time series with up to 3 or 6-lagged target series, and taking the naive approach with up to 3 or 6-lagged target series were, respectively, 0.25, 0.05, 0.27, 0.24, 0.29 and 0.26. The same scenarios for LightGBM resulted in 0.26, 0.04, 0.27, 0.26, 0.25, and 0.21.

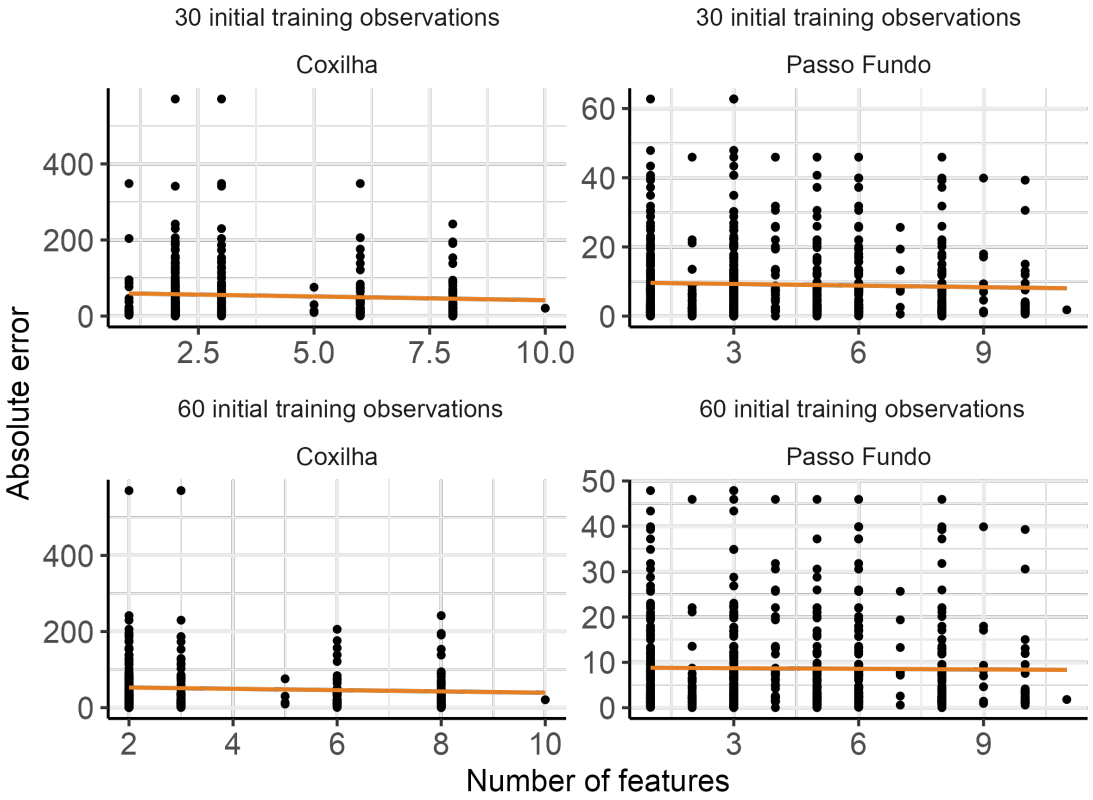

Figure 3.12: Scatter plots of the Random Forests' forecasting absolute error and the number of selected features of $\mathcal{D}$ (the sum of the number of climate, target time series and their lags) per forecast by our approach for the datasets of Coxilha and Passo Fundo regions.

Figure 3.12 shows that for Coxilha, the obtained correlation between the number of selected features of $\mathcal{D}$ (the sum of the number of climate, target time series and their lags) based on the reconstruction approach and the absolute error of each prediction based on Random Forests for each dataset for 30 and 60 initial training samples are, respectively, -0.06 and -0.06. For Passo Fundo, the obtained correlation between both variables for 30 and 60 initial training samples are -0.04 and -0.01.

3.2.3.2 Simulation study

Table 3.4 presents the estimated parameters of the autoregressive model with exogenous time series described in Equation 3.11 and their standard errors for the simulation study. Figure 3.13 illustrates the time series generated based on the Poisson and negative binomial ARX by presenting a sample of one time series per case of the simulation study.

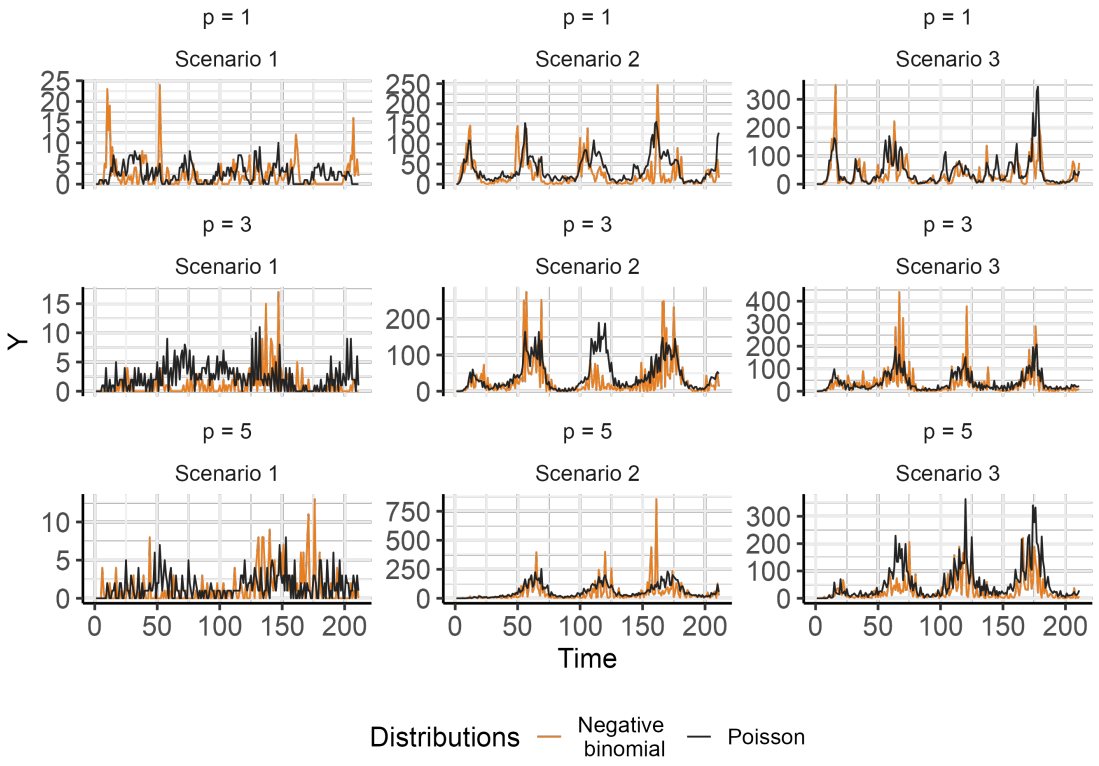

Figure 3.13: A sample of one simulated time series built upon Poisson and negative binomial ARX considering the lags $p = \{1, 3, 5\}$ and the scenarios: 1 - No influence of climate time series presented in Table 3.3 on simulated insect abundances; 2 - influence of five climate time series presented in Table 3.3 on simulated insect abundances; and 3 - Influence of all climate time series presented in Table 3.3 on simulated insect abundances.

Considering all scenarios, approaches and learning algorithms, the obtained average RMSE for the Poisson ARX was 24.3 with a standard deviation (sd) of 19.40 for training the learning algorithms with 30 initial samples and an average of 25.1

(sd = 20.70) for 60 initial samples. The obtained average RMSE for the negative binomial ARX was 39.1 (sd = 30.80) for training the learning algorithms with 30 initial samples and an average of 40.2 (sd = 32.40) for 60 initial samples. Given the slight difference between initial training samples on the forecasting performances (Pearson correlation and RSME), we solely present in Figure 3.14, Figure 3.15, Figure 3.16 and Figure 3.17 the performance of the learning algorithms trained with 30 initial training samples used to start the one-step ahead forecasting for Poisson and negative binomial ARX.

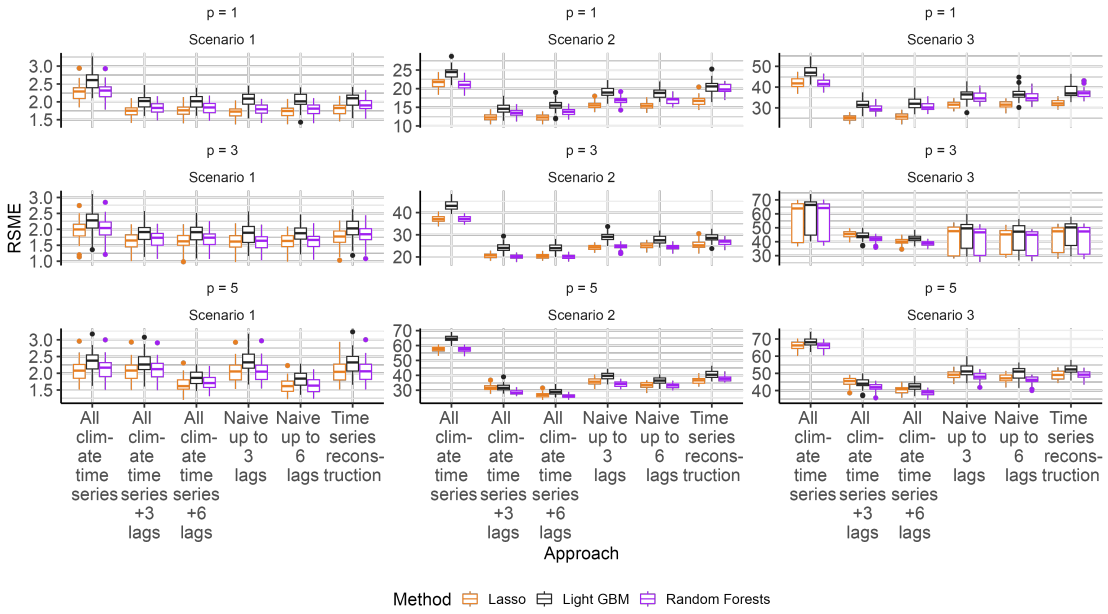

Figure 3.14: Root Mean Squared Error (RMSE) metric obtained by the Random Forests, Lasso-regularised linear regression, and LightGBM algorithms for each approach (including all climate time series with no lags, time series reconstruction, taking the naive approach with up to 3 or 6-step lagged target series, and all climate times series with up to 3 or 6-step lagged target series) considering the simulation study where the initial number of training samples is 30 and insect abundance is generated based on the **Poisson** ARX.

In Figure 3.14, the performance of the learning algorithms for the simulation study is presented based on the Poisson ARX. Overall, the average RMSE for Random Forests, Lasso-regularised linear regression and LightGBM were, respectively, 23.7 (sd = 18.80), 23.5 (sd = 19.00) and 25.8 (sd = 20.20), indicating that for the ma-

jority of the scenarios, values of p and approaches the learning algorithms have similar performances. However, some scenarios, such as Scenario 2, Lasso-regularised linear regression and Random Forests, perform better than LightGBM.

For scenario 1, where there is no influence of climate time series on the generation of insect abundances based on Poisson ARX, the average RMSE considering all values of p for the proposed method, all climate time series with no lags, all climate time series with up to 3 or 6-lagged target series, and taking the naive approach with up to 3 or 6-lagged target series were, respectively, 1.97 (sd = 0.30), 2.22 (sd = 0.34), 1.91 (sd = 0.32), 1.78 (sd = 0.25), 1.89 (sd = 0.34) and 1.75 (sd = 0.25). For scenario 2, where there is an influence of five climate time series based on Poisson ARX, the average RMSE considering all values of p for the proposed method, all climate time series with no lags, all climate time series with up to 3 or 6-lagged target series, and taking the naive approach with up to 3 or 6-lagged target series were, respectively, 28.1 (sd = 8.26), 40.5 (sd = 15.70), 21.9 (sd = 7.34), 20.9 (sd = 5.88), 26.5 (sd = 8.28) and 25.7 (sd = 7.38). For scenario 3, where there is an influence of all climate time series based on Poisson ARX, the average RMSE considering all values of p for the proposed method, all climate time series with no lags, all climate time series with up to 3 or 6-lagged target series, and taking the naive approach with up to 3 or 6-lagged target series were, respectively, 43.3 (sd = 8.31), 56.2 (sd = 12.2), 38.7 (sd = 7.67), 36.8 (sd = 5.93), 42.3 (sd = 8.81) and 41.5 (sd = 7.96).

In Figure 3.15, the performance of the learning algorithms for the simulation study is presented based on the negative binomial ARX. Overall, the average RMSE for Random Forests, Lasso-regularised linear regression and LightGBM were, respectively, 37.2 (sd = 28.90), 38.5 (sd = 30.60) and 41.7 (sd = 32.40), indicating that for all scenarios, values of p and approaches the learning algorithms have similar performances. For scenario 1, where there is no influence of climate time series on the generation of insect abundances based on negative binomial ARX, the average RMSE considering all lags for the proposed method, all climate time series with no lags, all climate time series with up to 3 or 6-lagged target series, and taking the naive approach with up to 3 or 6-lagged target series were, respectively, 2.60 (sd = 0.70), 2.75 (sd = 0.80), 2.51 (sd = 0.70), 2.43 (sd = 0.69), 2.49 (sd = 0.70)

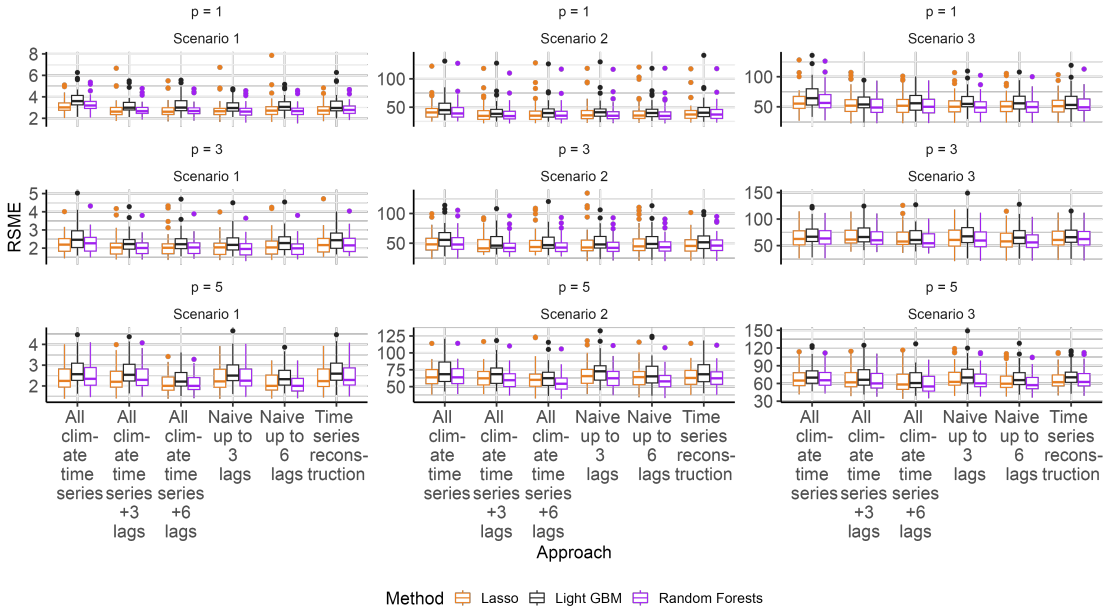

Figure 3.15: Root Mean Squared Error (RMSE) metric obtained by the Random Forests, Lasso-regularised linear regression, and LightGBM regression algorithms for each approach (including all climate time series with no lags, time series reconstruction, taking the naive approach with up to 3 or 6-step lagged target series, and all climate times with up to 3 or 6-step lagged target series) considering the simulation study where the initial number of training samples is 30 and insect abundance is generated based on the **negative binomial** ARX.

and 2.44 (sd = 0.72).

For scenario 2, where there is an influence of five climate time series based on negative binomial ARX, the average RMSE considering all lags for the proposed method, all climate time series with no lags, all climate time series with up to 3 or 6-lagged target series, and taking the naive approach with up to 3 or 6-lagged target series were, respectively, 53.0 (sd = 19.60), 56.0 (sd = 19.80), 50.9 (sd = 19.00), 50.2 (sd = 19.30), 53.1 (sd = 21.20) and 52.1 (sd = 20.20). For scenario 3, where there is an influence of all climate time series based on negative binomial ARX, the average RMSE considering all lags for the proposed method, all climate time series with no lags, all climate time series with up to 3 or 6-lagged target series, and taking the naive approach with up to 3 or 6-lagged target series were, respectively, 62.4 (sd = 19.30), 66.0 (sd = 19.70), 62.6 (sd = 19.30), 59.9

(sd = 18.50), 62.9 (sd = 20.70) and 60.4 (sd = 19.10).

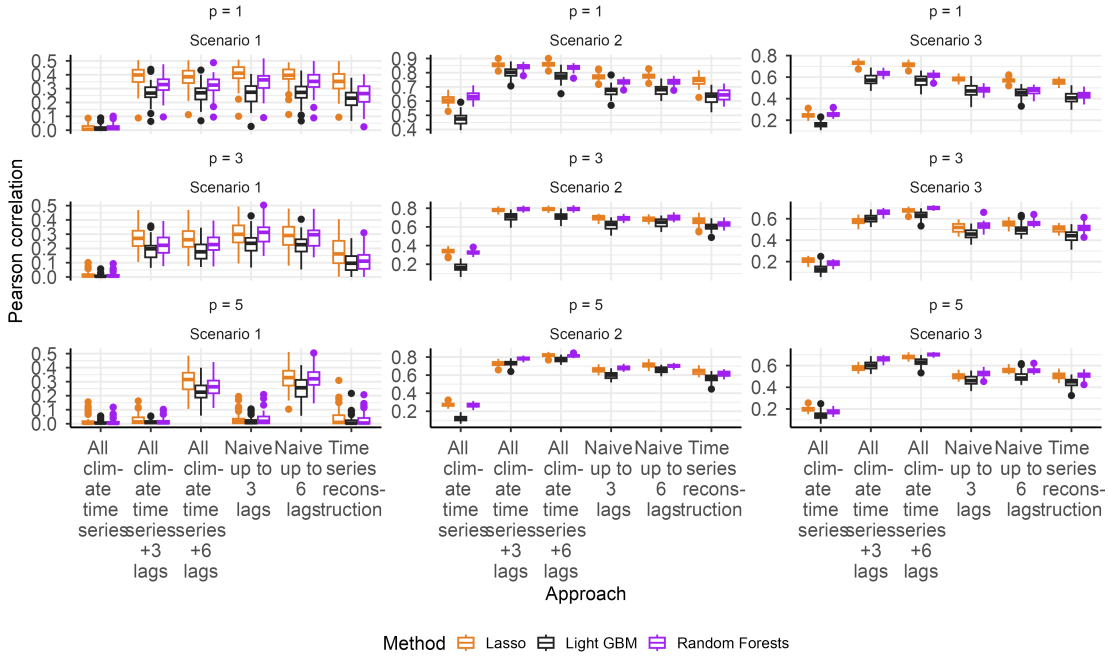

Figure 3.16: Pearson correlation metric obtained by the Random Forests, Lasso-regularised linear regression, and LightGBM algorithms for each approach (including all climate time series with no lags, time series reconstruction, taking the naive approach with up to 3 or 6-step lagged target series, and all climate times series with up to 3 or 6-step lagged target series) considering the simulation study where the initial number of training samples is 30 and insect abundance is generated based on the **Poisson** ARX.

In Figure 3.16, the correlation metric obtained by the learning algorithms for the simulation study is presented based on the Poisson ARX. Overall, the average correlation for Random Forests, Lasso-regularised linear regression and LightGBM were, respectively, 0.45 (sd = 0.25), 0.47 (sd = 0.25) and 0.40 (sd = 0.24), indicating that for all scenarios, values of p and approaches the Lasso-regularised linear regression and Random Forests algorithms have higher average correlation. For scenario 1, where there is no influence of climate time series on the generation of insect abundances based on negative binomial ARX, the average RMSE considering all lags for the proposed method, all climate time series with no lags, all climate time series with up to 3 or 6-lagged target series, and taking the naive

approach with up to 3 or 6-lagged target series were, respectively, 0.14 (sd = 0.13), 0.01 (sd = 0.02), 0.19 (sd = 0.15), 0.27 (sd = 0.09), 0.21 (sd = 0.16) and 0.30 (sd = 0.01).

For scenario 2, where there is an influence of five climate time series based on Poisson ARX, the average correlation considering all lags for the proposed method, all climate time series with no lags, all climate time series with up to 3 or 6-lagged target series, and taking the naive approach with up to 3 or 6-lagged target series were, respectively, 0.64 (sd = 0.06), 0.35 (sd = 0.17), 0.78 (sd = 0.05), 0.80 (sd = 0.05), 0.68 (sd = 0.06) and 0.70 (sd = 0.05). For scenario 3, where there is an influence of all climate time series based on negative binomial ARX, the average correlation considering all lags for the proposed method, all climate time series with no lags, all climate time series with up to 3 or 6-lagged target series, and taking the naive approach with up to 3 or 6-lagged target series were, respectively, 0.48 (sd = 0.06), 0.19 (sd = 0.04), 0.62 (sd = 0.06), 0.66 (sd = 0.05), 0.50 (sd = 0.06) and 0.52 (sd = 0.06).

In Figure 3.17, the correlation metric obtained by the learning algorithms for the simulation study is presented based on the negative binomial ARX. Overall, the average correlation for Random Forests, Lasso-regularised linear regression and LightGBM were, respectively, 0.17 (sd = 0.12), 0.15 (sd = 0.12) and 0.13 (sd = 0.10), indicating that for all scenarios, values of p and approaches the Lasso-regularised linear regression and Random Forests algorithms have higher average correlation. For scenario 1, where there is no influence of climate time series on the generation of insect abundances based on negative binomial ARX, the average RMSE considering all lags for the proposed method, all climate time series with no lags, all climate time series with up to 3 or 6-lagged target series, and taking the naive approach with up to 3 or 6-lagged target series were, respectively, 0.06 (sd = 0.08), 0.01 (sd = 0.02), 0.10 (sd = 0.10), 0.14 (sd = 0.08), 0.12 (sd = 0.10) and 0.17 (sd = 0.08).

For scenario 2, where there is an influence of five climate time series based on negative binomial ARX, the average correlation considering all lags for the proposed method, all climate time series with no lags, all climate time series with up to 3

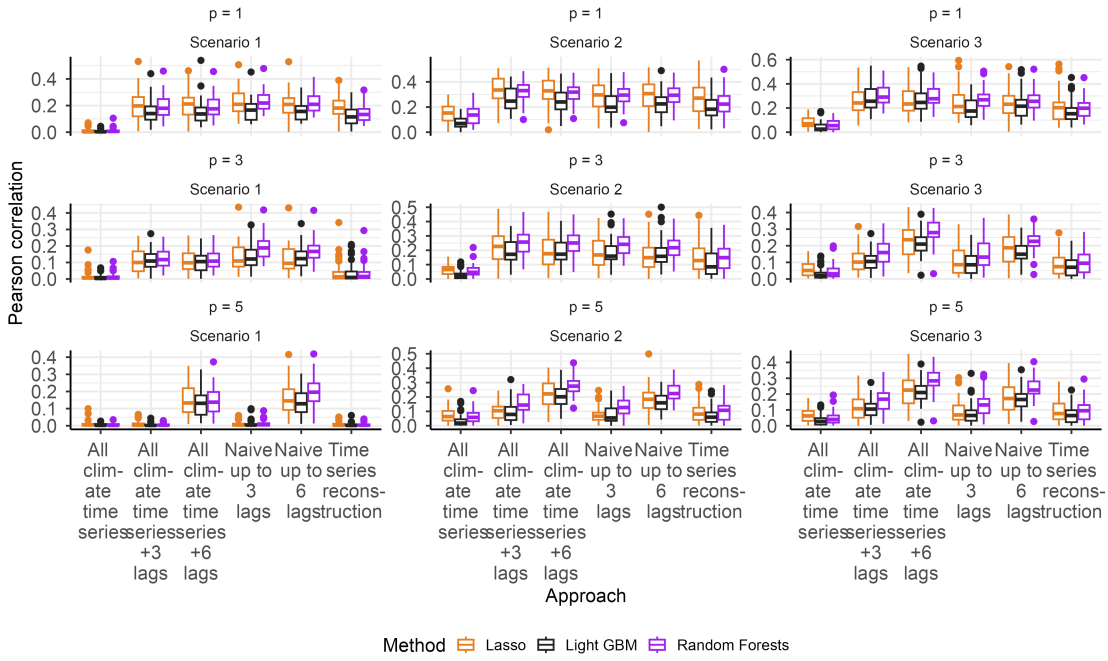

Figure 3.17: Pearson correlation metric obtained by the Random Forests, Lasso-regularised linear regression, and LightGBM algorithms for each approach (including all climate time series with no lags, time series reconstruction, taking the naive approach with up to 3 or 6-step lagged target series, and all climate times series with up to 3 or 6-step lagged target series) considering the simulation study where the initial number of training samples is 30 and insect abundance is generated based on the **negative binomial** ARX.

or 6-lagged target series, and taking the naive approach with up to 3 or 6-lagged target series were, respectively, 0.15 (sd = 0.11), 0.07 (sd = 0.06), 0.21 (sd = 0.12), 0.25 (sd = 0.10), 0.19 (sd = 0.11) and 0.21 (sd = 0.10). For scenario 3, where there is an influence of all climate time series based on negative binomial ARX, the average correlation considering all lags for the proposed method, all climate time series with no lags, all climate time series with up to 3 or 6-lagged target series, and taking the naive approach with up to 3 or 6-lagged target series were, respectively, 0.12 (sd = 0.09), 0.05 (sd = 0.04), 0.18 (sd = 0.11), 0.25 (sd = 0.09), 0.15 (sd = 0.11) and 0.21 (sd = 0.09).

Figure 3.18 shows the average correlation between the number of selected features of $\mathcal{D}$ (the sum of the number of climate, target time series and their lags) and

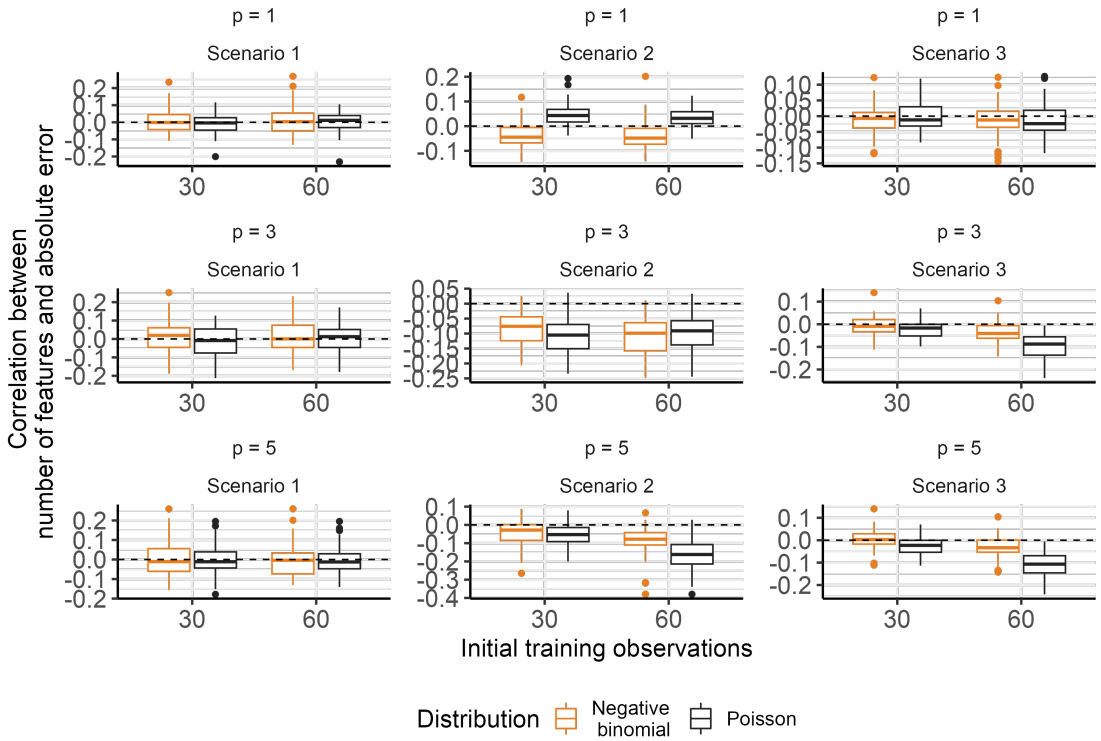

Figure 3.18: Boxplots of the correlation between the Random Forests's forecasting absolute error and the number of selected features of $\mathcal{D}$ (the sum of the number of climate, target time series and their lags) per forecast by our approach for the simulated study. The correlations are presented for the case where the Random Forests algorithm was trained with 30 and 60 initial observation to start the one-step ahead forecasting. The dashed line indicates the correlation equal to zero.

the absolute prediction error using our approach with Random Forests to obtain forecasts. Overall, the percentage of negative correlations considering all scenarios and values of p based on the Poisson ARX for 30 and 60 initial samples to start the one-step ahead forecasting were 63.9% and 69.0%. Based on the negative binomial ARX for 30 and 60, initial samples to start the one-step ahead forecasting were 61.6% and 71.3%. Also, the average correlations considering all scenarios and values of p based on the Poisson ARX for 30 and 60 initial samples to start the one-step ahead forecasting were -0.02 (sd $= 0.07$) and -0.05 (sd $= 0.09$). Based on the negative binomial ARX for 30 and 60 initial samples to start the one-step ahead forecasting were -0.02 (sd $= 0.07$) and -0.03 (sd $= 0.08$). It indicates that

the correlation is close to zero, and most correlations are negative.

For scenario 1, the percentage of negative correlations considering all values of p based on the Poisson ARX for 30 and 60 initial samples to start the one-step ahead forecasting were 55.5% and 48.6%. Based on the negative binomial ARX for 30 and 60 initial samples to start the one-step ahead forecasting were 49.3% and 50.7%. Also, the average correlations considering all values of p based on the Poisson ARX for 30 and 60 initial samples to start the one-step ahead forecasting were -0.01 (sd = 0.07) and 0.00 (sd = 0.07). Based on the negative binomial ARX for 30 and 60 initial samples to start the one-step ahead forecasting were 0.01 (sd = 0.08) and 0.01 (sd = 0.08).

For scenario 2, the percentage of negative correlations considering all values of p based on the Poisson ARX for 30 and 60 initial samples to start the one-step ahead forecasting were 66.0% and 70.8%. Based on the negative binomial ARX for 30 and 60 initial samples to start the one-step ahead forecasting were 83.3% and 91.3%. In addition, the average correlations considering all values of p based on the Poisson ARX for 30 and 60 initial samples to start the one-step ahead forecasting were -0.04 (sd = 0.08) and -0.08 (sd = 0.10). Based on the negative binomial ARX for 30 and 60 initial samples to start the one-step ahead forecasting were -0.05 (sd = 0.06) and -0.08 (sd = 0.08).

For scenario 3, the percentage of negative correlations considering all values of p based on the Poisson ARX for 30 and 60 initial samples to start the one-step ahead forecasting were 70.0% and 87.5%. Based on the negative binomial ARX for 30 and 60 initial samples to start the one-step ahead forecasting were 52.0% and 72.0%. The average correlations considering all values of p based on the Poisson ARX for 30 and 60 initial samples to start the one-step ahead forecasting were -0.01 (sd = 0.04) and -0.07 (sd = 0.07). Based on the negative binomial ARX for 30 and 60 initial samples to start the one-step ahead forecasting were 0.00 (sd = 0.05) and -0.03 (sd = 0.05).

3.2.4 Discussion

We presented a new approach for reconstructing time series dependencies using Takens' embedding theorem and Granger's causality. The approach can automatically select target and climate time series, including their lags, and we propose using machine learning algorithms that "learn" from the reconstructed time series to forecast insect abundance. We applied our proposed methods to two different datasets of insect time series and climate covariates associated with every observation of insect abundance. Also, a simulation study was introduced to explore the novel approach. The case study illustrates that the proposed approach is competitive compared with the other approaches presented in this chapter, regardless of the performance metric used. Also, the correlations presented by Figure 3.12 bring insights concerning the effect of the number of selected features on the forecasting performance, indicating that more features do not negatively impact the overall performance. The approaches using a combination of climate and the target time series performed better than those using solely the climate time series. It indicates the importance of using the target time series as features for the machine learning algorithms for forecasting insect abundance.

The selection of the presented performance metrics aids the understanding of the behaviour of the forecasts in a direction by looking at the Pearson correlation and proximity of the forecasts by analysing the RMSE (although other complementary performance metrics could be used). However, the presented ones allowed compare forecasts provided by the approaches [Koutsandreas et al., 2022]. In terms of RMSE performance, the simulation based on Poisson ARX emphasised the importance of adding the target time series as features for the machine learning methods. When the importance of the climate time series was highlighted in scenarios 2 and 3, the poor performance of the approach that uses all climate time series solely with no lags to predict insect abundance is more evident compared to the others. For the simulation study based on negative binomial ARX, most approaches obtained similar performances due to the super dispersion provided by the negative binomial distribution. It indicates that when the association of climate and the target time series are blurred by super dispersion, the implemented approaches perform similarly.

The correlation highlighted that by increasing the values of p, the approaches considering the lags 3 and 6 tend to have a higher correlation based on the Poisson ARX. This result is highlighted by Scenario 1. However, when we analyse the RMSE for these approaches, we can see that their average values are smaller than our proposed approach, which indicates that the direction of the insect dynamics can be better captured but not its absolute values. Moreover, the same conclusions can be applied to the simulation results related to the negative binomial ARX, mostly in Scenario 1, and for most of the other scenarios, the performances among the approaches were similar, showing the feasibility of our method. The other scenarios for both ARX models also indicate the feasibility of our method with competitive performances with these alternatives.

Figure 3.18 highlights the effect of the number of selected features of $\mathcal{D}$ (the sum of the number of climate, target time series and their lags) on the forecasting performance of the proposed approach based on the Random Forests' forecasting absolute error. Our results showed that most simulation scenarios' correlations are closer to zero or negative. We only observed negative and positive correlations with relatively higher intensity in scenario 2. It indicates the importance of selecting the lags and climate time series for forecasting insect abundances. Therefore, our approach provides selection criteria for lags and climate time series that impact the forecast performance of the Random Forests algorithm in the minority of the simulated scenarios. The impact is translated into selecting fewer climate time series in scenarios where they present an influence on the simulated insect abundance.

Our results indicate that the proposed method is competitive with the other approaches to applying machine learning to forecast insect abundances. Other researchers have proposed causal discovery methods for time series analysis based on different methodologies [Eichler and Didelez, 2010, Eichler, 2013, Runge et al., 2019b, Glymour et al., 2019, Assaad et al., 2022, Yuan and Shou, 2022, Runge et al., 2023]. Moreover, several authors have been exploring the application of deep learning models with attention mechanism [Vaswani, 2017] and variations of the transformer architectures to time series [Ahmed et al., 2023, Tong et al., 2023]. However, combining machine learning algorithms and causal discovery based on

Takens' embedding theory targeting forecasting presents a novel contribution to entomologists applying a learning-based algorithm to forecasting insect abundance. Thus, our work will be a basis for developing new techniques in this domain. Future work includes the investigation of potential improvements to our proposed approach by introducing deep learning mechanisms. We identified two main limitations in this work. The first relates to the focus on machine learning-based predictions, where the exploration of other classical time-series forecasting models was not presented as part of the comparison study. In future work, we will explore other classical methods to improve the proposed approach and expand the comparison space. The second point is the low correlation performance reported in the study, which shows the challenge of using the presented models to forecast insect abundances. As future work, we will explore ways to improve the performance of the proposed approach, focusing on alternative techniques for reconstructing time-series dependencies.

3.2.5 Conclusion

The proposed approach presented a competitive performance with commonly used approaches to forecast insect abundance in terms of predictive power. It shows the feasibility of applying this approach to forecasting insect pest abundance, and therefore, this study constitutes a basis for developing new techniques to predict insect outbreaks. Also, the feature selection of our method does not negatively influence the forecasting capability of the Random Forests algorithm, indicating that even if a climate time series does not influence the target time series, our methodology can select features that maintain a competitive forecasting performance compared to approaches presented in our study.

Scenarios	p	Intercept	ω_1	ω_2	ω_3	ω_4	ω_5	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6
	1	0.58 (0.10)	0.60 (0.06)	-	-	-	-	-	-	-	-	-	-
1	3	0.49 (0.10)	0.50 (0.06)	0.24 (0.07)	- 0.08 (0.07)	-	-	-	-	-	-	-	-
	5	0.51 (0.10)	0.51 (0.07)	0.23 (0.08)	- 0.08 (0.08)	0.04 (0.08)	−0.04 (0.07)	-	-	-	-	-	-
	1	0.14 (5.49)	0.59 (0.06)	-	-	-	-	−0.16 (0.19)	−0.01 (0.24)	0.18 (0.38)	−0.00 (0.00)	−0.00 (0.03)	-
2	3	−3.27 (4.46)	0.49 (0.07)	0.22 (0.08)	−0.11 (0.08)	-	-	0.15 (0.18)	0.01 (0.23)	−0.12 (0.36)	−0.00 (0.00)	0.04 (0.03)	-
	5	−1.55 (4.55)	0.43 (0.07)	0.22 (0.08)	−0.02 (0.08)	0.04 (0.08)	−0.04 (0.07)	0.01 (0.18)	0.03 (0.22)	0.02 (0.35)	−0.00 (0.00)	0.00 (0.03)	-
	1	5.61 (5.80)	0.59 (0.06)	-	-	-	-	−0.11 (0.20)	0.09 (0.26)	0.05 (0.41)	−0.00 (0.00)	−0.04 (0.03)	−0.00 (0.09)
3	3	−1.01 (4.86)	0.49 (0.07)	0.22 (0.08)	−0.13 (0.08)	-	-	−0.124 (0.18)	−0.04 (0.23)	−0.09 (0.36)	−0.00 (0.00)	0.05 (0.03)	0.04 (0.09)
	5	2.46 (5.19)	0.43 (0.08)	0.24 (0.08)	−0.02 (0.09)	−0.01 (0.08)	−0.044 (0.08)	−0.03 (0.18)	−0.02 (0.23)	0.01 (0.36)	−0.00 (0.00)	−0.01 (0.03)	0.06 (0.08)

Scenarios	p	θ_7	θ_8	θ_9	θ_{10}	θ_{11}	θ_{12}	θ_{13}	θ_{14}	θ_{15}	θ_{16}	θ_{17}	θ_{18}
	1	-	-	-	-	-	-	-	-	-	-	-	-
1	3	-	-	-	-	-	-	-	-	-	-	-	-
	5	-	-	-	-	-	-	-	-	-	-	-	-
	1	-	-	-	−0.01 (0.19)	−0.16 (0.25)	0.26 (0.37)	−0.00 (0.00)	0.01 (0.03)	-	-	-	-
2	3	-	-	-	0.01 (0.18)	0.00 (0.22)	0.06 (0.35)	−0.00 (0.00)	−0.01 (0.03)	-	-	-	-
	5	-	-	-	0.02 (0.17)	−0.08 (0.22)	0.13 (0.35)	−0.00 (0.00)	0.00 (0.03)	-	-	-	-
	1	−0.26 (0.22)	−0.09 (0.09)	0.00 (0.10)	0.07 (0.19)	−0.06 (0.25)	0.23 (0.40)	−0.00 (0.00)	0.01 (0.03)	−0.19 (0.09)	0.00 (0.22)	0.05 (0.09)	−0.14 (0.12)
3	3	−0.21 (0.21)	0.00 (0.08)	0.07 (0.10)	0.07 (0.19)	0.12 (0.23)	0.02 (0.36)	0.00 (0.00)	−0.03 (0.03)	−0.14 (0.08)	−0.02 (0.20)	0.02 (0.08)	−0.14 (0.10)
	5	−0.26 (0.20)	−0.03 (0.08)	0.15 (0.10)	0.09 (0.18)	0.07 (0.23)	0.02 (0.35)	0.00 (0.00)	−0.02 (0.03)	−0.15 (0.08)	0.05 (0.20)	0.01 (0.08)	−0.14 (0.10)

Table 3.4: Estimated ARX parameters and standard errors (parenthesis) for scenario 1 (no influence of climate time series on simulated insect abundances), scenario 2 (influence of five climate time series on simulated insect abundances), and scenario 3 (influence of all climate time series on simulated insect abundances) with lags, $p = \{1, 3, 5\}$.

CHAPTER 4

Conclusions

In this thesis, we have introduced a set of approaches using machine vision and statistical machine learning to improve the state-of-the-art quantitative methods applied to challenges on two of Caughley's management actions: 1) carry out monitoring programs without additional actions over stable populations and 2) decrease excessive populations. Initially, we use preprocessing techniques, such as RGB modifications of spectrograms and images, to classify avian and insect species based on sound and images in monitoring systems. We also provided a pipeline to combining computer vision and machine learning methods applied to entomology and illustrated this pipeline with a multi-class classification problem with small and imbalanced dataset challenges. Finally, we presented a new learning algorithm focused on predicting animal outbreaks based solely on abundance patterns, and we developed a new approach that can be applied to any learning algorithm when focused on time series and prediction of animal abundances.

In Chapter 2, we explored the use of deep learning, machine vision, machine learning and feature engineering methods to detect and identify animal species through three studies. In the first section, we investigated the feasibility of using pre-trained VGG16 architecture for avian species detection through sound spectrograms. The obtained validation accuracy demonstrated the viability of this approach in detecting the studied avian species. Our results indicated that using coloured im-

ages to represent spectrograms generalises classification better than grey-scale and histogram-equalised images. Considering the challenge presented by a multiclass classification, the high accuracy obtained in the study shows promising results for further exploration of using coloured spectrograms for avian identification. These findings can serve as the basis for developing future animal monitoring programs based on sound recording of avian species, which can significantly enhance sampling efforts without increasing costs. In the second section of Chapter 2, we presented an overview of machine vision methods for identifying and localising insects in laboratory and field settings. We discussed machine learning, computer vision methods, and relevant software and hardware considerations related to computing platforms. Given the rapidly evolving nature of machine vision, we aimed to provide readers with a range of currently available options for identification and localisation tasks while emphasising that methodology selection depends on the specific research question and target species. Finally, we proposed a pipeline integrating computer vision and machine learning methods that can be extended to multiple entomological applications. The Pairwise Controlled Manifold Approximation Projection (PaCMAP) algorithm also incorporates non-supervised learning techniques to evaluate the feature space in vision problems, highlighting their use in entomology.

The third section of Chapter 2 presented a novel approach to classify species of the *Anastrepha pseudoparallela* group using classical computer science techniques combined with machine learning and computer vision methods. The primary challenges addressed in this work relate to the group's rare species characteristics, resulting in imbalanced and scarce data. We demonstrated that combining proposed features collected from the wing's node structure with RGB information from specific polygons defined by specialists outperformed the traditional VGG16-PCA architecture across various learning algorithms. In addition, our results showed that utilising coloured images with SMOTE and Random Forests algorithms provided superior classification performance for these species. This study serves as an initial exploration of the classification of rare *Anastrepha* species. It provides a foundation for classifying fruit fly species from other groups that face similar data scarcity and class imbalance challenges. In Chapter 3, we developed statis-

tical modelling approaches for animal monitoring and control. The first section introduced the Pattern-based Prediction (PBP) method and analysed its sensitivity and performance through simulation studies and actual data application. We applied the method to a time series of aphids in wheat crops in Southern Brazil. By building upon the alert zone procedure framework, we enhanced the information extractable from the population dynamics of species of interest. This process allows for extracting different population states using dynamics obtained from monitoring programs. In addition, by grouping these states into different cluster matrices, we observed the frequency of each pattern type occurring before outbreaks, facilitating outbreak classification based on different pattern types. The PBP method allows further exploration by incorporating covariates to improve outbreak predictions. In the second section of Chapter 3, we proposed a feature selection approach for forecasting insect abundance that demonstrated competitive performance compared to commonly used approaches in terms of MSE and Pearson correlation. Our method proved feasible for forecasting insect pest abundance, establishing a basis for developing new techniques to predict insect outbreaks. The feature selection component of our method did not negatively influence the forecasting capability of the Random Forests algorithm, indicating that even when climate time series do not directly impact the target time series, our methodology can select features that maintain competitive forecasting performance compared to other approaches presented in the study. The proposed approaches presented in Chapter 2 and Chapter 3 introduce innovative methods directly applicable to animal monitoring and Caughley's management actions.

Overall, for the first Caughley's management actions, carrying out monitoring programs without additional actions over stable populations, we conclude that the use of transfer learning through VGG16 architecture combined with the proposed machine vision pipeline and feature engineering helped to explore and present approaches with higher classification performance among the presented applications. In future research, we will explore the combination of explainable artificial intelligence algorithms to the presented techniques to aid the evaluation of the features used by future proposed approaches. In addition, we will combine various data sources, including audio, image, and morphometric features, to aid in animal

classification. Finally, regarding Caughley's management actions on decreasing excessive populations, we conclude that the introduced learning algorithm for insect outbreak classification and the proposed approach for insect abundance forecasting demonstrated competitive performance based on the presented benchmark. Also, their design enabled us to investigate these algorithms with greater transparency. For the PBP, the patterns and indices can be analysed to understand a particular prediction better. For the proposed approach to insect abundance forecasting, the feature selection procedure introduces more clarity in investigating the effect of exogenous time series on a specific forecast. In the future, we aim to incorporate additional exogenous time series into the PBP algorithm, targeting an increase in classification performance. Moreover, for the proposed approach, we will incorporate deep learning techniques, such as the use of time series attention mechanisms, to introduce more nonlinear interactions among exogenous time series.

The central focus of this thesis was the development of new methodologies aimed at enhancing automatic animal monitoring systems. While we did not aim to create a fully integrated end-to-end system that combines all the proposed methods and approaches, the outlined proposals can be integrated into a framework for fully automatic monitoring systems of animal populations, particularly insect pests. These methods have significant implications for ecological research and agricultural pest management. Future studies could explore integrating these approaches into a unified system, which would greatly benefit ecologists and stakeholders in the field by automating and refining animal monitoring processes. All implementations and analyses presented in this thesis are designed to be reproducible, with methodologies available to interested practitioners. The code for the machine learning models, statistical analyses, and forecasting frameworks is completely documented to encourage wider use of the proposed approaches in ecological monitoring and pest management applications.

Appendix

4.A Pattern-Based Prediction of Population Outbreaks

We introduce the `pypbp` package, which is a Python implementation of the Pattern-Based Prediction (PBP) method, and examples of how to use it in Section 4.A.1. Also, we present the clustering algorithm used by the PBP method in Section 4.A.2. Finally, we describe the methods used for the PBP optimisation procedure, including a comparison between the performance of Generalised Simulated Annealing (GSA) and differential evolution in Section 4.A.3.

4.A.1 The `pypbp` package

To use our package, first install Python 3 using the official website (https://www.python.org). We recommend the Jupyter Lab environment (https://jupyter.org) as a GUI for Python, however there are many other options. Using the `pip` command in your terminal (`cmd` in Windows, or `terminal` in Mac and Linux operating systems), you may install Jupyter Lab by executing

```
pip install jupyterlab
```

To install the `pypbp` package, execute

```
pip install pypbp
```

You may then open a Jupyter Lab environment using the command

```
jupyter lab --core-mode
```

Finally, you may create a new Jupyter notebook file using the Jupyter Lab environment. Using the cells of the Jupyter notebook file, you can import and use functions implemented in the `pypbp` package, as presented below. More information is available in the package description page (https://pypbp-documentation.readthedocs.io/en/latest/) and GitHub repository (https://github.com/GabrielRPalma/PyPBP).

4.A.1.1 A minimal reproducible example

The following code obtains estimates for the hyperparameters in the PBP method based on the aphid data used as motivation in the chapter.

```
import pypbp as pbp

results, xstar, clustered_patterns, parameters = pbp.pbp_fit(
    time_series = pbp.time_series,
              train_percentage = 0.5,
              xstar = 200,
              maxfun = 1000)
pbp_plot(time_series = pbp.time_series, clustered_patterns =
    clustered_patterns,
               parameters = parameters,
               xnew = [10, 60, 20, 10, 20, 80, 90])
```

By default, the optimisation is performed using 5-fold cross-validation by carrying out the GSA method with the negative area under the ROC curve as the objective function to be minimised, without pre-processing the time series using empirical mode decomposition. The function `pbp_fit` contains other arguments, such as `verbose`, which prints the area under the ROC curve at every iteration of the optimisation process, and `maxfun`, which sets the number of evaluations of the objective function by the optimisation algorithm (GSA as default).

The object `clustered_patterns` contains the cluster matrices estimated using the time series presented, the object `results` contains a data frame with the metrics obtained from the model: accuracy, fI-Score, precision, recall, true positive rate (TPR), false positive rate (FPR) and the estimated area under the ROC curve. It also contains the estimates for m, d^*_{cluster} and α. These metrics are presented

for each of the four criteria used in our methodology to select d^*_{base} (maximum FPR of 0.1 and 0.2 and minimum TPR of 0.8 and 0.9). The `parameters` object contains the estimates for m, d^*_{cluster} and α in a dictionary ready for use. The function `pbp_plot` can be used to visualise the obtained patterns. Finally, the details of each function used in the `pypbp` package are presented on the website https://pypbp-documentation.readthedocs.io/en/latest/.

4.A.2 Pattern clustering algorithm

Algorithm 4.A.2 starts with pattern $\mathbf{p}_1$, which represents the first row of the matrix $\mathbf{P}$. We remove $\mathbf{p}_1$ from $\mathbf{P}$ and add it as the first row of $\mathbf{P}'_1$. After that we compute the association metric between $\mathbf{p}_1$ and all subsequent rows of $\mathbf{P}$. If $d(\mathbf{p}_1, \mathbf{p}_j) \geq d^*_{\text{cluster}}$, we add pattern $\mathbf{p}_j$ as the last row of the cluster matrix $\mathbf{P}'_1$ and delete it from $\mathbf{P}$. We repeat this process to obtain the cluster matrices $\mathbf{P}'_c, c = 1, \ldots, C \leq I$, until there are no more rows left in $\mathbf{P}$.

Algorithm 3 Obtaining cluster matrices $\mathbf{P}'_c, c = 1, \ldots, C$, from $\mathbf{P}$.

Input: d^*_{cluster} and $\mathbf{P}$
set $c = 1$
set $l =$ number of rows of $\mathbf{P}$
for i **in** $\{1, 2, \ldots, l\}$ **do**
  add $\mathbf{p}_i$ as the first row of $\mathbf{P}'_c$ and delete it from $\mathbf{P}$
  set $l' =$ number of rows of $\mathbf{P}'_c$
  for j **in** $\{1, 2, \ldots, (I - \sum_c l'_c)\}$ **do**
    if $d(\mathbf{p}_i, \mathbf{p}_j) \geq d^*_{\text{cluster}}$ **then**
      append $\mathbf{p}_j$ as the last row of $\mathbf{P}'_c$ and delete $\mathbf{p}_j$ from $\mathbf{P}$
    end if
    update $l' =$ number of rows of $\mathbf{P}'_c$
  end for
  update $c = c + 1$
  update $l =$ number of remaining rows of $\mathbf{P}$
end for
Output: cluster matrices $\mathbf{P}'_c, c = 1, \ldots, C \leq I$

4.A.3 Comparison between optimisation methods

Here, we compare the differential evolution [Storn and Price, 1997] and the method [Tsallis and Stariolo, 1996] algorithms used to estimate α, d^*_{cluster} and m. We use

these methods to optimise the area under the ROC curve (AUROC). We looked at the performance of our method using the time series data on aphids and parasitoids collected in Coxilha (Brazil-São Paulo). In Figure 4.A.1 we present the accuracy, in Figure 4.A.2 the area below the ROC curve (AUROC), in Figures 4.A.3 and 4.A.4 we present the true and false positive rates. In Figures 4.A.5, 4.A.6, and 4.A.7 we present the estimated α, m and d^*_{cluster} hyperparameters.

Differential evolution is an algorithm based on evolutionary computation, which briefly consists of setting a population of candidate solutions and creating new updated candidate solutions based on the existing based on the score. For more detail see Storn and Price [1997]. On the other hand, the generalised simulated annealing is a stochastic algorithm used for finding the global minimum of a given function in a continuous D-dimensional space. The method utilises a generalised entropic form

$$S_q = k\frac{1-\sum_i p_i^q}{q-1},$$

where $q \in \mathbb{R}$, p_i are probabilities of the microscopic configurations and k is a conventional positive constant [Tsallis and Stariolo, 1996]. Based on this entropy Tsallis and Stariolo [1996] generalised its formulation to include the cases of the Boltzmann Cauchy machines, allowing to find local minima according to an acceptance temperature set in the generalised metropolis algorithm. Differential evolution was approximately ten times slower than generalised method. This fact drove our decision to choose the algorithm as the default algorithm for optimising the hyperparameters involved in the PBP method. However, users are allowed to choose which method to use when utilising the `pypbp` package.

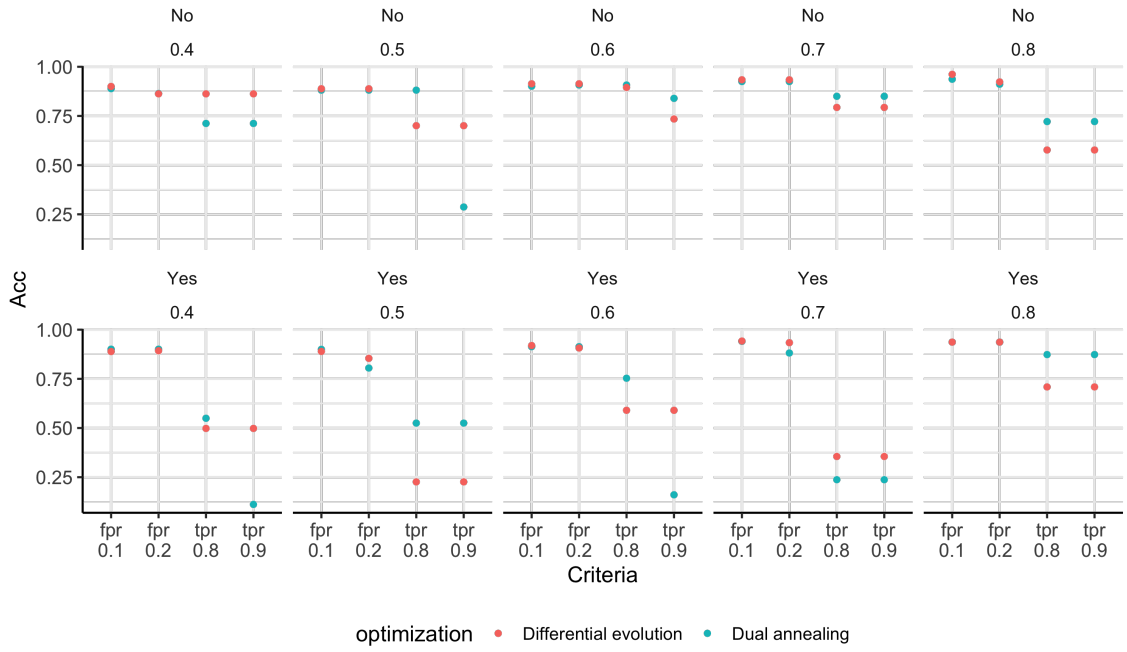

Figure 4.A.1: Accuracy obtained on the testing step of the PBP method. In both algorithms the following methods to choose d^*_{base} were used: based on minimum FPR (.1 and .2 as minimum) and TFR maximum (.8 and .9 as maximum).

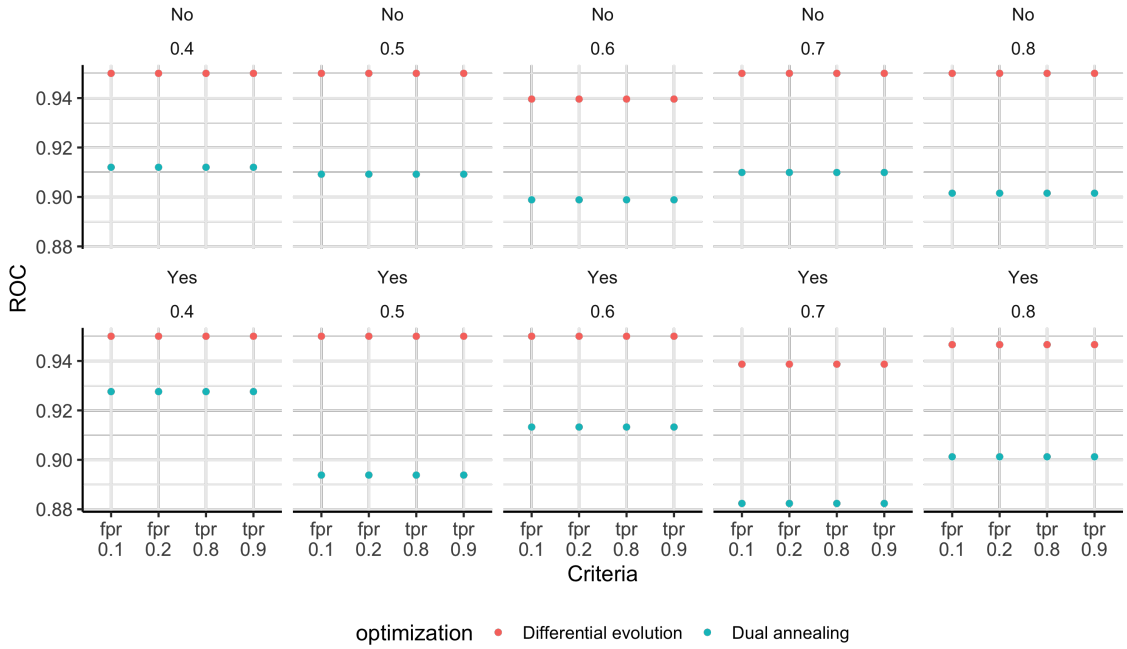

Figure 4.A.2: Area below the ROC curve obtained on the training step of the PBP method. In both algorithms the following methods to choose d^*_{base} were used: based on minimum FPR (0.1 and 0.2 as minimum) and TFR maximum (0.8 and 0.9 as maximum).

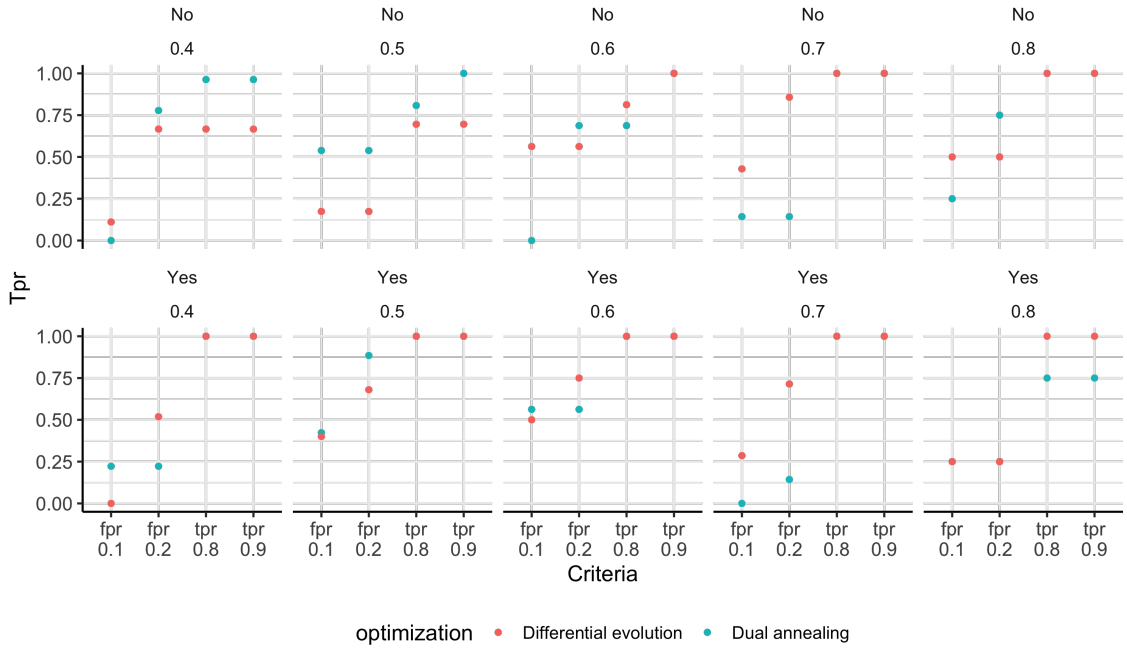

Figure 4.A.3: True Positive Rate obtained on the testing step of the PBP method. In both algorithms the following methods to choose d^*_{base} were used: based on minimum FPR (0.1 and 0.2 as minimum) and TFR maximum (0.8 and 0.9 as maximum).

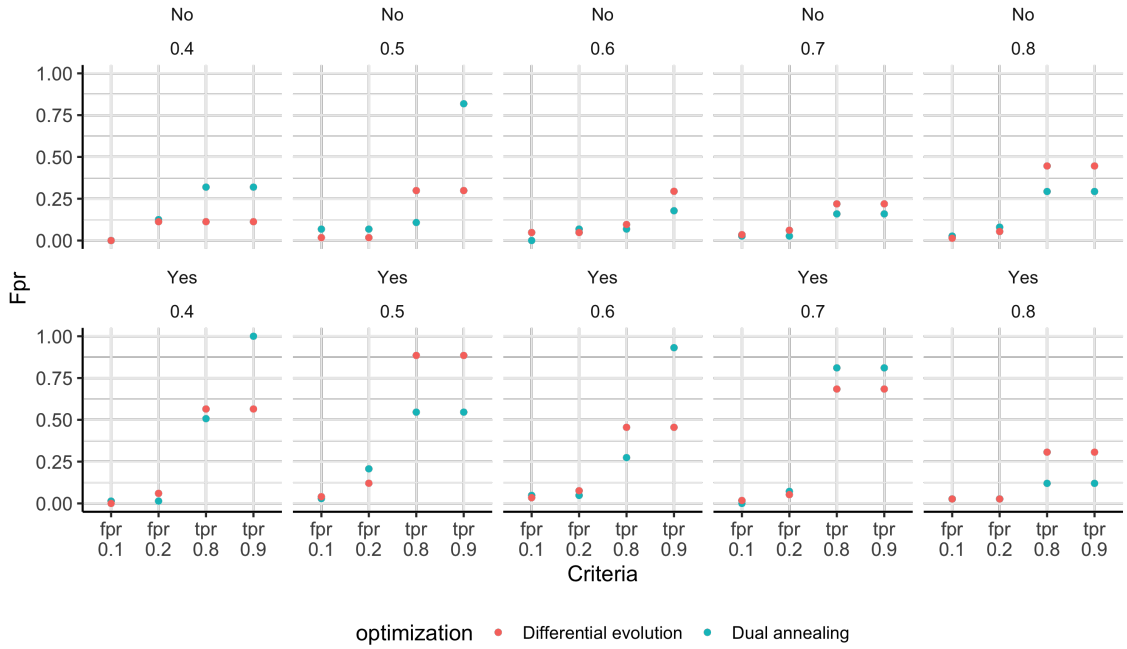

Figure 4.A.4: False Positive Rate obtained on the testing step of the PBP method. In both algorithms the following methods to choose d^*_{base} were used: based on minimum FPR (0.1 and 0.2 as minimum) and TFR maximum (0.8 and 0.9 as maximum).

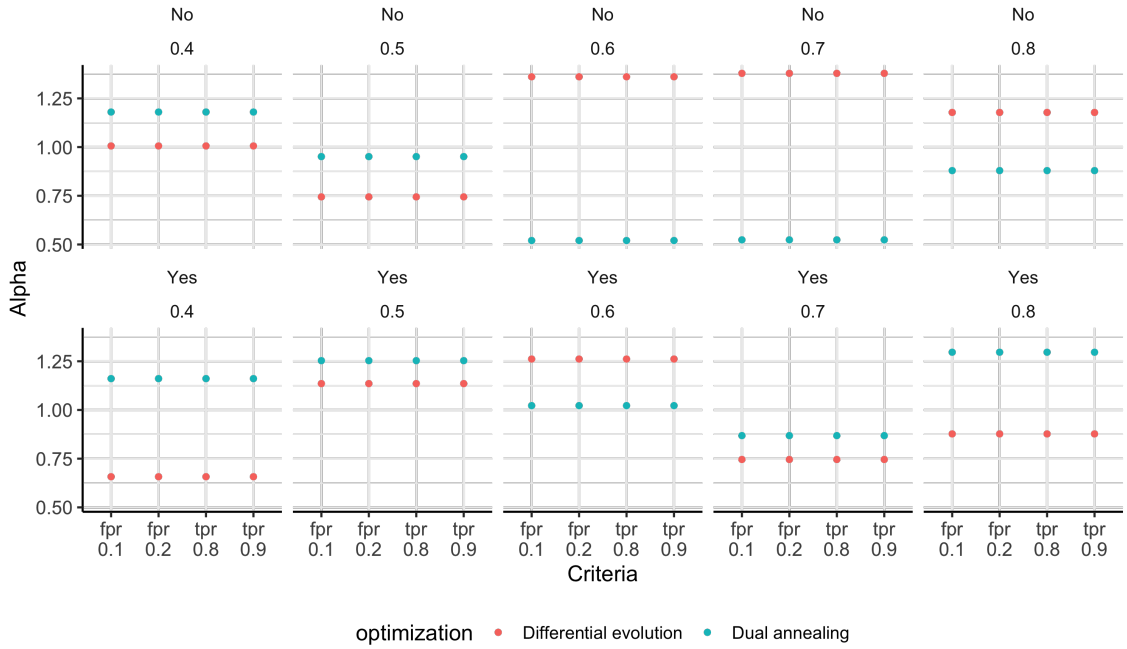

Figure 4.A.5: Estimated parameter α obtained on the training step of the PBP method. In both algorithms the following methods to choose d^*_{base} were used: based on minimum FPR (0.1 and 0.2 as minimum) and TFR maximum (0.8 and 0.9 as maximum).

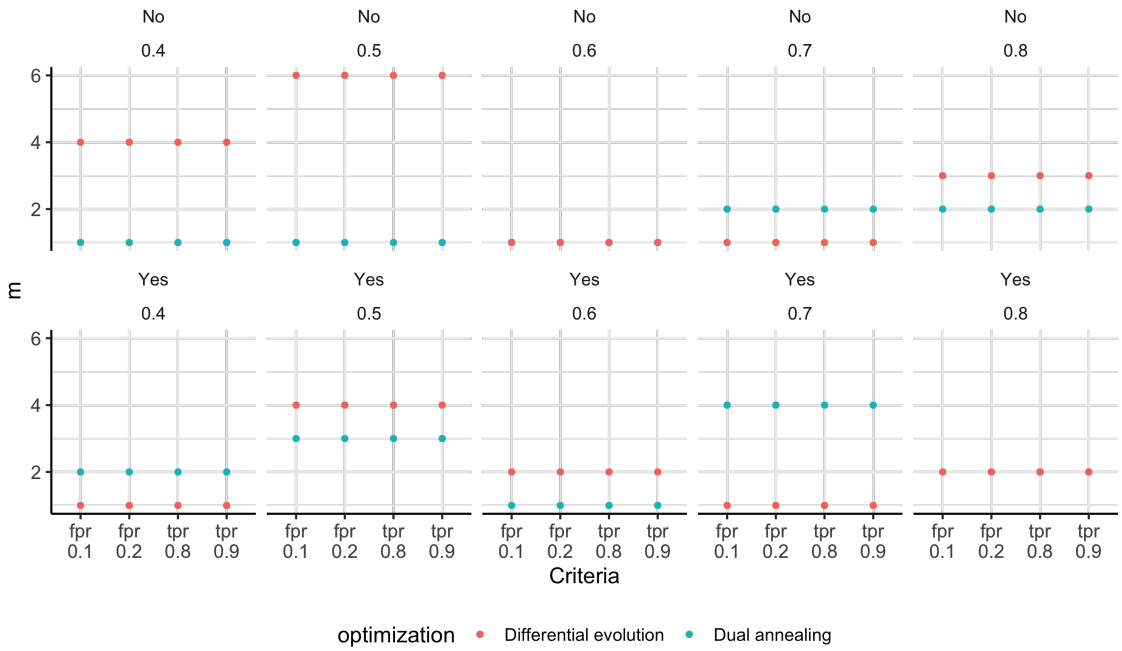

Figure 4.A.6: Estimated parameter m obtained on the training step of the PBP method. In both algorithms the following methods to choose d^*_{base} were used: based on minimum FPR (0.1 and 0.2 as minimum) and TFR maximum (0.8 and 0.9 as maximum).

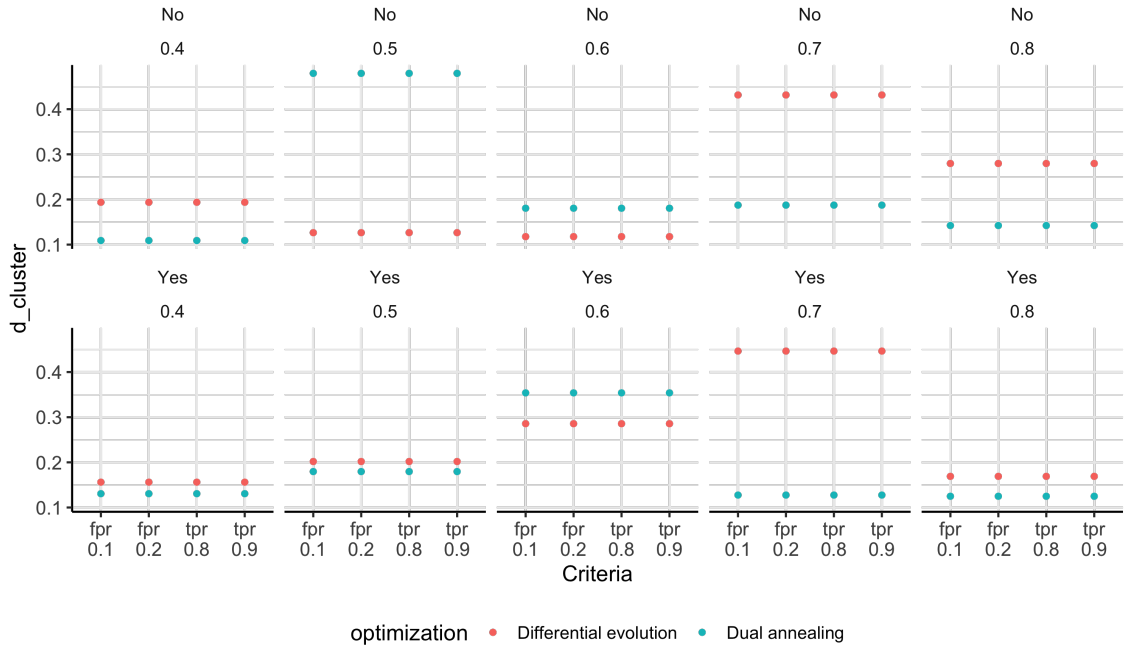

Figure 4.A.7: Estimated parameter d^*_{cluster} obtained on the training step of the PBP method. In both algorithms the following methods to choose d^*_{base} were used: based on minimum FPR (0.1 and 0.2 as minimum) and TFR maximum (0.8 and 0.9 as maximum).

Bibliography

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. Tensorflow: Large-scale machine learning on heterogeneous systems, 2015. URL https://www.tensorflow.org/. Software available from tensorflow.org. Accessed: 2025-01-08. 75

Penchala Abhinav and Sunitha Dhavale. Bird species classification and prediction using machine learning algorithms. In *2024 IEEE Pune Section International Conference (PuneCon)*, pages 1 – 5. IEEE, 2024. 7

Nilanjana Adhikari, Suman Bhattacharya, and Mahamuda Sultana. A modified resnet152v2 framework for bird species classification. *Innovations in Systems and Software Engineering*, pages 1 – 14, 2024. 7

Sabeen Ahmed, Ian E Nielsen, Aakash Tripathi, Shamoon Siddiqui, Ravi P Ramachandran, and Ghulam Rasool. Transformers in time-series analysis: A tutorial. *Circuits, Systems, and Signal Processing*, 42:7433 – 7466, 2023. 129

T Mitchell Aide, Carlos Corrada-Bravo, Marconi Campos-Cerqueira, Carlos Milan, Giovany Vega, and Rafael Alvarez. Real-time bioacoustics monitoring and automated species identification. *PeerJ*, 1:e103, 2013. 6

Eman A. Al-Shahari, Ghadah Aldehim, Nabil Sharaf Almalki, Mohammed Assiri, Ahmed Sayed, and Mrim M. Alnfiai. Innovative insect detection and classification for the agricultural sector using gannet optimization algorithm with deep learning. *IEEE Access*, 12:108041 – 108051, 2024. 67

Slade Allen-Ankins, Sebastian Hoefer, Jacopo Bartholomew, Sheryn Brodie, and Lin Schwarzkopf. The use of birdnet embeddings as a fast solution to find novel sound classes in audio recordings. *Frontiers in Ecology and Evolution*, 12:12 – 1409407, 2025. 6

Josep Alós, Kim Aarestrup, David Abecasis, Pedro Afonso, Alexandre Alonso-Fernandez, Eneko Aspillaga, Margarida Barcelo-Serra, Jonathan Bolland, Miguel Cabanellas-Reboredo, Robert Lennox, et al. Toward a decade of ocean science for sustainable development through acoustic animal tracking. *Global Change Biology*, 28:5630 – 5653, 2022. 2

Ahmed Alqaraawi, Martin Schuessler, Philipp Weiß, Enrico Costanza, and Nadia Berthouze. Evaluating saliency map explanations for convolutional neural networks: a user study. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 275 – 285, 2020. 51

Abderraouf Amrani, Dean Diepeveen, David Murray, Michael GK Jones, and Ferdous Sohel. Multi-task learning model for agricultural pest detection from crop-plant imagery: A bayesian approach. *Computers and Electronics in Agriculture*, 218, 2024. 66

Dimitrios Androutsos, KN Plataniotiss, and Anastasios N Venetsanopoulos. Distance measures for color image retrieval. In *Proceedings 1998 International Conference on Image Processing. ICIP98 (Cat. No. 98CB36269)*, volume 2, pages 770 – 774. IEEE, 1998. 89

Alexandre S Araújo, Allen L Norrbom, Roberto A Zucchi, and Marcoandre Savaris. A new species of the anastrepha pseudoparallela group (diptera: Tephritidae) with a synopsis of the group in Brazil. *Neotropical Entomology*, 53:854 – 867, 2024. 14, 67, 68

J Arthy, K Raja, Vignesh Raghuraman, S Ashwath, and Rahul Bundele. Avian species detection using spectogram and CNN. In *The Role of Artificial Intelligence in Advancing Applied Life Sciences*, pages 265 – 290. IGI Global Scientific Publishing, 2025. 7

Keyvan Asefpour Vakilian and Jafar Massah. Performance evaluation of a machine vision system for insect pests identification of field crops using artificial neural networks. *Archives of Phytopathology and plant protection*, 46:1262 – 1269, 2013. 25

Charles K Assaad, Emilie Devijver, and Eric Gaussier. Entropy-based discovery of summary causal graphs in time series. *Entropy*, 24, 2022. 129

Mohammed Assiri, Elmouez Samir Abd Elhameed, Arun Kumar, and Chinu Singla. Automated insect detection and classification using pelican optimization algorithm with deep learning on internet of enabled agricultural sector. *SN Computer Science*, 5, 2024. 66

Rohil Badkundri, Victor Valbuena, Srikusmanjali Pinnamareddy, Brittney Cantrell, and Janet Standeven. Forecasting the 2017-2018 yemen cholera outbreak with machine learning. *arXiv preprint*, 02 2019. 85

S Balasubramaniam, C Vijesh Joe, A Prasanth, and K Satheesh Kumar. Computer vision systems in livestock farming, poultry farming, and fish farming: Applications, use cases, and research directions. *Computer Vision in Smart Agriculture and Crop Management*, pages 221 – 258, 2025. 3

JS Bale, JC Van Lenteren, and F Bigler. Biological control and sustainable food production. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363:761 – 776, 2008. 24

Ms Aayushi Bansal, Dr Rewa Sharma, and Dr Mamta Kathuria. A systematic review on data scarcity problem in deep learning: solution and applications. *ACM Computing Surveys (Csur)*, 54:1 – 29, 2022. 6

Kiran Gandhi Bapatla, Ajay Deep Singh, Vennila Sengottaiyan, Rajasekhara Rao Korada, and Srujana Yeddula. Impact of climate change on helicoverpa armigera

voltinism in different agro-climatic zones of india. *Journal of Thermal Biology*, 106:103229, 2022. 8

José A Barbero-Aparicio, Alicia Olivares-Gil, Juan J Rodríguez, César García-Osorio, and José F Díez-Pastor. Addressing data scarcity in protein fitness landscape analysis: A study on semi-supervised and deep transfer learning techniques. *Information Fusion*, 102, 2024. 66

Marco Barzman, Paolo Bàrberi, A Nicholas E Birch, Piet Boonekamp, Silke Dachbrodt-Saaydeh, Benno Graf, Bernd Hommel, Jens Erik Jensen, Jozsef Kiss, Per Kudsk, et al. Eight principles of integrated pest management. *Agronomy for sustainable development*, 35:1199 – 1215, 2015. 7

James R Bell, Lynda Alderson, Daniela Izera, Tracey Kruger, Sue Parker, Jon Pickup, Chris R Shortall, Mark S Taylor, Paul Verrier, and Richard Harrington. Long-term phenological trends, species accumulation rates, aphid traits and climate: Five decades of change in migrating aphids. *Journal of Animal Ecology*, 84:21 – 34, 2015. 97

Youssef Benseddik, Abdelmalek Boutaleb Joutei, Abdelali Blenzar, Said Amiri, Adil Asfers, Fouad Mokrini, and Rachid Lahlali. Biological control potential of moroccan entomopathogenic nematodes for managing the flatheaded rootborer, capnodis tenebrionis (linné) (coleoptera: Buprestidae). *Crop Protection*, 158:105991, 2022. ISSN 0261-2194. doi: https://doi.org/10.1016/j.cropro.2022.105991. URL https://www.sciencedirect.com/science/article/pii/S0261219422000874. 7

Marc Besson, Jamie Alison, Kim Bjerge, Thomas E Gorochowski, Toke T Høye, Tommaso Jucker, Hjalte MR Mann, and Christopher F Clements. Towards the fully automated monitoring of ecological communities. *Ecology Letters*, 25:2753 – 2775, 2022. 1

Kim Bjerge, Quentin Geissmann, Jamie Alison, Hjalte MR Mann, Toke T Høye, Mads Dyrmann, and Henrik Karstoft. Hierarchical classification of insects with multitask learning and anomaly detection. *Ecological Informatics*, 77, 2023. 67

Kim Bjerge, Henrik Karstoft, Hjalte M.R. Mann, and Toke T. Høye. A deep learning pipeline for time-lapse camera monitoring of insects and their floral environments. *Ecological Informatics*, 84, 2024. ISSN 1574-9541. doi: https://doi.org/10.1016/j.ecoinf.2024.102861. URL https://www.sciencedirect.com/science/article/pii/S1574954124004035. 5

Jarrett D Blair, Kaitlyn M Gaynor, Meredith S Palmer, and Katie E Marshall. A gentle introduction to computer vision-based specimen classification in ecological datasets. *Journal of Animal Ecology*, 93:147 – 158, 2024. 4, 5

Marek L Borowiec, Rebecca B Dikow, Paul B Frandsen, Alexander McKeeken, Gabriele Valentini, and Alexander E White. Deep learning as a tool for ecology and evolution. *Methods in Ecology and Evolution*, 13:1640 – 1660, 2022. 67

Yan Boulanger, Adèle Desaint, Véronique Martel, Maryse Marchand, Salomon Massoda Tonye, Rémi Saint-Amant, and Jacques Régnière. Recent climate change strongly impacted the population dynamic of a north american insect pest species. *PLOS Climate*, 4:e0000488, 2025. 8

George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015. 109

Marek Brabec, Alois Honěk, Stano Pekár, and Zdenka Martinková. Population dynamics of aphids on cereals: digging in the time-series data to reveal population regulation caused by temperature. *PloS one*, 9, 2014. 108

G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000. Accessed: 2025-01-08. 69

Jelto Branding, Dieter von Hörsten, Elias Böckmann, Jens Karl Wegener, and Eberhard Hartung. Insectsound1000 an insect sound dataset for deep learning based acoustic insect recognition. *Scientific Data*, 11, 2024. 66

Francisco J Bravo Sanchez, Md Rahat Hossain, Nathan B English, and Steven T Moore. Bioacoustic classification of avian calls from raw sound waveforms with an open-source deep learning architecture. *Scientific Reports*, 11:1 – 12, 2021. 16

Matthew W Breece, Matthew J Oliver, Dewayne A Fox, Edward A Hale, Danielle E Haulsee, Matthew Shatley, Steven J Bograd, Elliott L Hazen, and Heather Welch. A satellite-based mobile warning system to reduce interactions with an endangered species. *Ecological Applications*, 31:e02358, 2021. 2

Leo Breiman. Random Forests. *Mach. Learn.*, 45(1):5–32, October 2001. ISSN 0885-6125. doi: 10.1023/A:1010933404324. URL https://doi.org/10.1023/A:1010933404324. 98

Lorenzo Brigato and Luca Iocchi. A close look at deep learning with small data. In *2020 25th international conference on pattern recognition (ICPR)*, pages 2490 – 2497. IEEE, 2021. 66

Benjamin C Bright, Andrew T Hudak, Arjan JH Meddens, Joel M Egan, and Carl L Jorgensen. Mapping multiple insect outbreaks across large regions annually using landsat time series data. *Remote Sensing*, 12, 2020. 84

M Briseño-Jaramillo, A Hutschenreiter, F Aureli, and AA Ríos-Chelén. Passive acoustic monitoring and acoustic indices reveal noise-related changes in bird vocalisations. *Bioacoustics*, 34:167 – 187, 2025. 6

Cheryl Brown, Lori Lynch, and David Zilberman. The economics of controlling insect-transmitted plant diseases. *American Journal of Agricultural Economics*, 84:279 – 291, 2002. 106

Chris Brunsdon and Alexis Comber. Big issues for big data: challenges for critical spatial data analytics. *arXiv preprint arXiv:2007.11281*, 2020. 15

David Buckeridge. Outbreak detection through automated surveillance: A review of the determinants of detection. *Journal of biomedical informatics*, 40:370 – 9, 09 2007. doi: 10.1016/j.jbi.2006.09.003. 84, 85

TR Buckley. Charting a future for entomological taxonomy in new zealand. *New Zealand Entomologist*, pages 1 – 17, 2024. 3, 4

Regiane Cristina Bueno, Adeney de Freitas Bueno, Flávio Moscardi, José Roberto Parra, and Clara Hoffmann-Campo. Lepidopteran larva consumption of soybean foliage: Basis for developing multiple-species economic thresholds for pest

management decisions. *Pest management science*, 67:170 – 4, 10 2010. doi: 10.1002/ps.2047. 8, 86

Ulf Büntgen, Andrew Liebhold, Daniel Nievergelt, Beat Wermelinger, Alain Roques, Frederick Reinig, Paul J Krusic, Alma Piermattei, Simon Egli, Paolo Cherubini, et al. Return of the moth: rethinking the effect of climate on insect outbreaks. *Oecologia*, 192:543 – 552, 2020. 84

Joseph B Burant, Candace Park, Gustavo S Betini, and D Ryan Norris. Early warning indicators of population collapse in a seasonal environment. *Journal of Animal Ecology*, 90:1538 – 1549, 2021. 8

Howard Burkom, Wayne Loschen, Richard Wojcik, Rekha Holtry, Monika Punjabi, Martina Siwek, Sheri Lewis, et al. Electronic surveillance system for the early notification of community-based epidemics (essence): Overview, components, and public health applications. *JMIR public health and surveillance*, 7, 2021. 84

Daniel Büttner and Markus Rabe. Sales forecasting in the electrical industry-an illustrative comparison of time series and machine learning approaches. In *2021 9th International Conference on Traffic and Logistic Engineering (ICTLE)*, pages 69 – 78. IEEE, 2021. 107

Jennifer Byrne, Robert Lillywhite, Henry Creissen, Fiona Thorne, and Lael Walsh. Quantifying integrated pest management adoption in food horticulture. *Crop Protection*, page 107165, 2025. 7

Joel Cabrera and Edwin Villanueva. Investigating generative neural-network models for building pest insect detectors in sticky trap images for the peruvian horticulture. In *Annual International Conference on Information Management and Big Data*, pages 356 – 369. Springer, 2021. 67

Carolina Tieppo Camarozano, Aloísio Coelho Jr, Ranyse Barbosa Querino da Silva, and José Roberto Postali Parra. Can trichogramma atopovirilia oatman & platner replaces trichogramma galloi zucchi for diatraea saccharalis (fabricius) control? *Scientia Agricola*, 79, 2021. 24, 25

Bruna Campos Paula, Lilian Luchesi, and Patrícia Monticelli. Railway noise and long-distance calls of free-living maned wolves in ecological station of itirapina, são paulo, Brazil. *The Journal of the Acoustical Society of America*, 151:A147–A147, 2022. doi: 10.1121/10.0010927. URL https://doi.org/10.1121/10.0010927. 16

Sarah Carmesin, David Woller, David Parker, Miroslav Kulich, and Masoumeh Mansouri. The hamiltonian cycle and travelling salesperson problems with traversal-dependent edge deletion. *Journal of Computational Science*, 74, 2023. 70

Graeme Caughley. Directions in conservation biology. *Journal of animal ecology*, pages 215 – 244, 1994. xi, 1, 2

Ana Ceia-Hasse, Carla A Sousa, Bruna R Gouveia, and César Capinha. Forecasting the abundance of disease vectors with deep learning. *Ecological Informatics*, 78, 2023. 107

Ta-Chien Chan, Jia-Hong Tang, Cheng-Yu Hsieh, Kevin J Chen, Tsan-Hua Yu, and Yu-Ting Tsai. Approaching precision public health by automated syndromic surveillance in communities. *Plos one*, 16, 2021. 85

Elissa M. Chasen, Dan J. Undersander, and Eileen M. Cullen. Revisiting the economic injury level and economic threshold model for potato leafhopper (hemiptera: Cicadellidae) in alfalfa. *Journal of Economic Entomology*, 108: 1748–1756, 05 2015. ISSN 0022-0493. doi: 10.1093/jee/tov120. URL https://doi.org/10.1093/jee/tov120. 8

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321 – 357, 2002. 71, 73

Shi Chen, Ari Whiteman, Ang Li, Tyler Rapp, Eric Delmelle, Gang Chen, Cheryl L Brown, Patrick Robinson, Maren J Coffman, Daniel Janies, et al. An operational machine learning approach to predict mosquito abundance based on socioeconomic and landscape patterns. *Landscape Ecology*, 34:1295 – 1311, 2019. 107, 115

Yanping Chen, Adena Why, Gustavo Batista, Agenor Mafra-Neto, and Eamonn Keogh. Flying insect classification with inexpensive sensors. *Journal of insect behavior*, 27:657 – 677, 2014. 66

Zekai Cheng, Rongqing Huang, Rong Qian, Wei Dong, Jingbo Zhu, and Meifang Liu. A lightweight crop pest detection method based on convolutional neural networks. *Applied Sciences*, 12, 2022. 26

François Chollet. *Deep Learning with Python*. Manning Publications, Shelter Island, NY, 2nd edition, 2018. ISBN 978-1617294433. 18, 44

Chih-Hsun Chou, Chang-Hsing Lee, and Hui-Wen Ni. Bird species recognition by comparing the hmms of the syllables. In *Proceedings of the Second International Conference on Innovative Computing, Information and Control (ICICIC)*, pages 143–143, 10 2007. ISBN 0-7695-2882-1. doi: 10.1109/ICICIC.2007.199. 16

Andrew Chow and Mamoudou Sétamou. Parasitism of diaphorina citri (hemiptera: Liviidae) by tamarixia radiata (hymenoptera: Eulophidae) on residential citrus in texas: Importance of colony size and instar composition. *Biological Control*, 165:104796, 2022. 24

Shawan Chowdhury, Michael D Jennions, Myron P Zalucki, Martine Maron, James EM Watson, and Richard A Fuller. Protected areas and the future of insect conservation. *Trends in Ecology & Evolution*, 38:85 – 95, 2023. 65

Sylvain Christin, Eric Hervet, and Nicolas Lecomte. Applications for deep learning in ecology. *Methods in Ecology and Evolution*, pages 1632–1644, 05 2018. doi: 10.1101/334854. 16

Physilia YS Chua, Sarah J Bourlat, Cameron Ferguson, Petra Korlevic, Leia Zhao, Torbjørn Ekrem, Rudolf Meier, and Mara KN Lawniczak. Future of DNA-based insect monitoring. *Trends in Genetics*, 39:531 – 544, 2023. 66

Catherine Matilda Collins, Hélène Audusseau, Chris Hassall, Nusha Keyghobadi, Palatty Allesh Sinu, and Manu E Saunders. Insect ecology and conservation in urban areas: An overview of knowledge and needs. *Insect conservation and diversity*, 17:169 – 181, 2024. 3

Fernando LuíS Cônsoli and José Roberto Postali Parra. Biology of trichogramma galloi and t. pretiosum (hymenoptera: Trichogrammatidae) reared in vitro and in vivo. *Annals of the Entomological Society of America*, 89:828 – 834, 1996. 24

Corinna Cortes and Vladimir Vapnik. Support-Vector Networks. *Mach. Learn.*, 20 (3):273–297, September 1995. ISSN 0885-6125. doi: 10.1023/A:1022627411411. URL https://doi.org/10.1023/A:1022627411411. 98

Koby Crammer and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of machine learning research*, 2 (Dec):265–292, 2001. 98

Jordan P Cuff and Allan Watt. Advances in insect biomonitoring for agriculture and forestry: A synthesis on a multifaceted special issue of agricultural and forest entomology. *Agricultural and Forest Entomology*, 27:1 – 7, 2025. 5, 6

Nabanita Das, Neelamadhab Padhy, Nilanjan Dey, Hrithik Paul, and Soumalya Chowdhury. Exploring explainable AI methods for bird sound-based species recognition systems. *Multimedia Tools and Applications*, 83:64223 – 64253, 2024. 6

Wang Dawei, Deng Limiao, Ni Jiangong, Gao Jiyue, Zhu Hongfei, and Han Zhongzhi. Recognition pest by image-based transfer learning. *Journal of the Science of Food and Agriculture*, 99:4524 – 4531, 2019. 25

Jean-Philippe Deguine, Jean-Noël Aubertot, Rica Joy Flor, Françoise Lescourret, Kris AG Wyckhuys, and Alain Ratnadass. Integrated pest management: good intentions, hard realities. a review. *Agronomy for Sustainable Development*, 41: 1 – 35, 2021a. 24

Jean-Philippe Deguine, Jean-Noël Aubertot, Rica Joy Flor, Françoise Lescourret, Kris AG Wyckhuys, and Alain Ratnadass. Integrated pest management: good intentions, hard realities. a review. *Agronomy for Sustainable Development*, 41: 38, 2021b. 7

Sheng Deng, Lan Du, Chen Li, Jun Ding, and Hongwei Liu. Sar automatic target recognition based on euclidean distance restricted autoencoder. *IEEE Journal*

of Selected Topics in Applied Earth Observations and Remote Sensing, 10:3323 – 3333, 2017. 74

Simone Disabato, Giuseppe Canonaco, Paul G Flikkema, Manuel Roveri, and Cesare Alippi. Birdsong detection at the edge with deep learning. In *2021 IEEE International Conference on Smart Computing (SMARTCOMP)*, pages 9 – 16. IEEE, 2021. 17

Koffi Djaman, Charles Higgins, Michael O'Neill, Shantel Begay, Komlan Koudahe, and Samuel Allen. Population dynamics of six major insect pests during multiple crop growing seasons in northwestern new mexico. *Insects*, 10, 2019. ISSN 2075-4450. doi: 10.3390/insects10110369. URL https://www.mdpi.com/2075-4450/10/11/369. 8

Thanh-Nghi Doan. Large-scale insect pest image classification. *Journal of Advances in Information Technology*, 14:328 – 341, 2023. 67

Tansel Dokeroglu, Ender Sevinc, Tayfun Kucukyilmaz, and Ahmet Cosar. A survey on new generation metaheuristic algorithms. *Computers & Industrial Engineering*, 137, 2019. ISSN 0360-8352. doi: https://doi.org/10.1016/j.cie.2019.106040. URL https://www.sciencedirect.com/science/article/pii/S0360835219304991. 71

Ganggang Dong, Guisheng Liao, Hongwei Liu, and Gangyao Kuang. A review of the autoencoder and its variants: A comparative perspective from target recognition in synthetic-aperture radar images. *IEEE Geoscience and Remote Sensing Magazine*, 6:44 – 68, 2018. 71

Marco Dorigo and Luca Maria Gambardella. Ant colony system: a cooperative learning approach to the traveling salesman problem. *IEEE Transactions on evolutionary computation*, 1:53 – 66, 1997. 71

Marco Dorigo, Mauro Birattari, and Thomas Stutzle. Ant colony optimization. *IEEE computational intelligence magazine*, 1:28 – 39, 2006. 71

Z. Dun, Paul Mitchell, and Mauro Agosti. Estimating diabrotica virgifera virgifera damage functions with field trial data: applying an unbalanced nested error

component model. *Journal of Applied Entomology*, 134:409 – 419, 12 2009. doi: 10.1111/j.1439-0418.2009.01487.x. 87

Rezvan Ehsani and Finn Drabløs. Robust distance measures for k nn classification of cancer data. *Cancer informatics*, 19, 2020. 89

Michael Eichler. Causal inference with multiple time series: principles and problems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371, 2013. 129

Michael Eichler and Vanessa Didelez. On granger causality and the effect of interventions in time series. *Lifetime data analysis*, 16:3 – 32, 2010. 129

Adam Ekholm, Ayco Tack, Pertti Pulkkinen, and Tomas Roslin. Host plant phenology, insect outbreaks and herbivore communities – the importance of timing. *Journal of Animal Ecology*, 11 2019. doi: 10.1111/1365-2656.13151. 86

Ahmed El-Ahmady, Medhat I Abul-Sood, Metwaly M Montaser, Ahmed M Galhom, and Ahmed Badry. Wing morphometric analysis of some species of the genus sarcophaga (diptera: Sarcophagidae) in egypt. *Egyptian Academic Journal of Biological Sciences. A, Entomology*, 17:79 – 81, 2024. 66

Isitor Emmanuel and Clare Stanier. Defining big data. In *Proceedings of the International Conference on Big Data and Advanced Wireless Technologies*, pages 1 – 6, 2016. 15

Eduardo Engel, Douglas Lau, Wesley AC Godoy, Mauricio PB Pasini, José B Malaquias, Carlos DR Santos, Juliana Pivato, and Paulo RV da S Pereira. Oscillation, synchrony, and multi-factor patterns between cereal aphids and parasitoid populations in southern Brazil. *Bulletin of Entomological Research*, 112: 143 – 150, 2022. 108, 112

Jean Étienne, Serge Quilici, Daniel Marival, and Antoine Franck. Biological control of diaphorina citri (hemiptera: Psyllidae) in guadeloupe by imported tamarixia radiata (hymenoptera: Eulophidae). *Fruits*, 56:307 – 315, 2001. 24

Syed Ruby Farah, S Giri Babu, and D Janardhan. Automated bird species identification using audio signal processing and convolutional neural network. In *2024 Second International Conference on Inventive Computing and Informatics (ICICI)*, pages 236 – 242. IEEE, 2024. 7

Colin Favret and Jeffrey M Sieracki. Machine vision automated species identification scaled towards production levels. *Systematic Entomology*, 41:133 – 143, 2016. 25

Jiedong Feng, Yaqin Sun, Kefei Zhang, Yindi Zhao, Yi Ren, Yu Chen, Huifu Zhuang, and Shuo Chen. Autonomous detection of spodoptera frugiperda by feeding symptoms directly from uav rgb imagery. *Applied Sciences*, 12, 2022. 26

Alberto Fernández, Salvador Garcia, Francisco Herrera, and Nitesh V Chawla. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of artificial intelligence research*, 61:863 – 905, 2018. 71, 73

Robert Finger. Digital innovations for sustainable and resilient agricultural systems. *European Review of Agricultural Economics*, 50:1277 – 1309, 06 2023. ISSN 0165-1587. doi: 10.1093/erae/jbad021. URL https://doi.org/10.1093/erae/jbad021. 2

Fateme Fotouhi, Kevin Menke, Aaron Prestholt, Ashish Gupta, Matthew E. Carroll, Hsin-Jung Yang, Edwin J. Skidmore, Matthew O'Neal, Nirav Merchant, Sajal K. Das, Peter Kyveryga, Baskar Ganapathysubramanian, Asheesh K. Singh, Arti Singh, and Soumik Sarkar. Persistent monitoring of insect-pests on sticky traps through hierarchical transfer learning and slicing-aided hyper inference. *Frontiers in Plant Science*, Volume 15 - 2024, 2024a. ISSN 1664-462X. doi: 10.3389/fpls.2024.1484587. URL https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2024.1484587. 4, 5

Fateme Fotouhi, Kevin Menke, Aaron Prestholt, Ashish Gupta, Matthew E Carroll, Hsin-Jung Yang, Edwin J Skidmore, Matthew O'Neal, Nirav Merchant, Sajal K Das, et al. Persistent monitoring of insect-pests on sticky traps through

hierarchical transfer learning and slicing-aided hyper inference. *Frontiers in Plant Science*, 15:1484587, 2024b. 5

Amy E Frazier and Lei Song. Artificial intelligence in landscape ecology: recent advances, perspectives, and opportunities. *Current Landscape Ecology Reports*, 10:1 – 13, 2025. 3

Frank Friedrich, Yoko Matsumura, Hans Pohl, Ming Bai, Thomas Hörnschemeyer, and Rolf G Beutel. Insect morphology in the age of phylogenomics: innovative techniques and its future role in systematics. *Entomological Science*, 17:1 – 24, 2014. 3, 4

John M Fryxell, Anthony RE Sinclair, and Graeme Caughley. *Wildlife ecology, conservation, and management*. John Wiley & Sons, 2014. 1

Lars Gamfeldt, Helmut Hillebrand, and Per R Jonsson. Multiple functions increase the importance of biodiversity for overall ecosystem functioning. *Ecology*, 89: 1223 – 1231, 2008. 65

Yuanyi Gao, Xiaobao Xue, Guoqing Qin, Kai Li, Jiahao Liu, Yulong Zhang, and Xinjiang Li. Application of machine learning in automatic image identification of insects-a review. *Ecological Informatics*, page 102539, 2024. 3, 66

Tiago Garcia, Luís Pina, Magnus Robb, Jorge Maria, Roel May, and Ricardo Oliveira. Long-range bird species identification using directional microphones and CNNs. *Machine Learning and Knowledge Extraction*, 6:2336 – 2354, 2024. ISSN 2504-4990. doi: 10.3390/make6040115. URL https://www.mdpi.com/2504-4990/6/4/115. 16

Amandine Gasc, Dante Francomano, John B Dunning, and Bryan C Pijanowski. Future directions for soundscape ecology: The importance of ornithological contributions. *The Auk: Ornithological Advances*, 134:215 – 228, 2017. 6, 15, 17

Kevin J Gaston. The magnitude of global insect species richness. *Conservation biology*, 5:283 – 296, 1991. 65

Pralhad Gavali and J Saira Banu. A novel approach to indian bird species identification: employing visual-acoustic fusion techniques for improved classification accuracy. *Frontiers in Artificial Intelligence*, 8:1527299, 2025. 6

Rémi Gerber, Christophe Piscart, Jean-Marc Roussel, and Benjamin Bergerot. Morphology-based classification of the flying capacities of aquatic insects: A first attempt. *Current zoology*, 70:607 – 617, 2024. 4

Rory Gibb, Ella Browning, Paul Glover-Kapfer, and Kate E Jones. Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring. *Methods in Ecology and Evolution*, 10:169 – 185, 2019. 6

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580 – 587, 2014. 26

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10, 2019. 129

H Charles J Godfray and HCJ Godfray. *Parasitoids: behavioral and evolutionary ecology*, volume 67. Princeton University Press, 1994. 86

Peter Goodell. Fifty years of the integrated control concept: The role of landscape ecology in ipm in san joaquin valley cotton. *Pest management science*, 65:1293 – 7, 12 2009. doi: 10.1002/ps.1859. 87

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016. 25, 26, 37, 39, 40, 44

Kaviraj Gosaye and Raj Kishen Moloo. A mobile application for fruit fly identification using deep transfer learning: A case study for mauritius. In *2022 International Conference for Advancement in Technology (ICONAT)*, pages 1 – 5. IEEE, 2022. 67

Kayleize Govender, Branden Ingram, and Pravesh Ranchod. Analysing the effect of latent space mutation strategies for pcgml. In *2024 IEEE Conference on Games (CoG)*, pages 1 – 8. IEEE, 2024. 73

Ivan Grijalva, Amanda R Skidmore, Marc A Milne, Paola Olaya-Arenas, Ian Kaplan, Rick E Foster, and John S Yaninek. Integrated pest management enhances biological control in a us midwestern agroecosystem by conserving predators and non-pest prey. *Agriculture, Ecosystems & Environment*, 368, 2024. 107

Gaurav Gupta, Meghana Kshirsagar, Ming Zhong, Shahrzad Gholami, and Juan Lavista Ferres. Comparing recurrent convolutional neural networks for large scale bird species classification. *Scientific reports*, 11:17085, 2021. 7, 16

Robert Guralnick, Raphael LaFrance, Michael Denslow, Samantha Blickhan, Mark Bouslog, Sean Miller, Jenn Yost, Jason Best, Deborah L Paul, Elizabeth Ellwood, et al. Humans in the loop: Community science and machine learning synergies for overcoming herbarium digitization bottlenecks. *Applications in Plant Sciences*, 12, 2024. 5, 6

Gebreegziabher Hailay Gebremariam. A systematic review of insect decline and discovery: Trends, drivers, and conservation strategies over the past two decades. *Psyche: A Journal of Entomology*, 2024, 2024. 65

Aidan Hall, Scott Johnson, James Cook, and Markus Riegler. High nymphal host density and mortality negatively impact parasitoid complex during an insect herbivore outbreak. *Insect Science*, 26, 08 2017. doi: 10.1111/1744-7917.12532. 86

Hala Hamdoun, Alaa Sagheer, and Hassan Youness. Energy time series forecasting-analytical and empirical assessment of conventional and machine learning models. *Journal of Intelligent & Fuzzy Systems*, 40:12477 – 12502, 2021. 107

Oskar LP Hansen, Jens-Christian Svenning, Kent Olsen, Steen Dupont, Beulah H Garner, Alexandros Iosifidis, Benjamin W Price, and Toke T Høye. Species-level image classification with convolutional neural network enables insect identification from habitus images. *Ecology and evolution*, 10:737 – 747, 2020. 26

Manfred Hartbauer. Artificial neuronal networks are revolutionizing entomological research. *Journal of Applied Entomology*, 148:232 – 251, 2024. 3, 4, 5

Florian Hartig, Nerea Abrego, Alex Bush, Jonathan M Chase, Gurutzeta Guillera-Arroita, Mathew A Leibold, Otso Ovaskainen, Loïc Pellissier, Maximilian Pichler, Giovanni Poggiato, et al. Novel community data in ecology-properties and prospects. *Trends in Ecology & Evolution*, 2024. 1

Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. The elements of statistical learning: Data mining, inference, and prediction. *Math. Intell.*, 27:83 – 85, 11 2004. doi: 10.1007/BF02985802. 25, 93

Trevor Hastie, Robert Tibshirani, Jerome Friedman, et al. The elements of statistical learning, 2009. 98

Hangfei He, Junyang Chen, Hongkun Chen, Borui Zeng, Yutong Huang, Yudan Zhaopeng, and Xiaoyan Chen. Enhancing insect sound classification using dual-tower network: A fusion of temporal and spectral feature perception. *Applied Sciences*, 14, 2024. 66

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770 – 778, 2016. 3

Mengqi He, Sanyi Tang, and Robert A Cheke. A holling type ii discrete switching host-parasitoid system with a nonlinear threshold policy for integrated pest management. *Discrete Dynamics in Nature and Society*, 2020:9425285, 2020. 7

Michelle Heck. Insect transmission of plant pathogens: A systems biology perspective. *MSystems*, 3:10 – 1128, 2018. 106

René Heinrich, Lukas Rauch, Bernhard Sick, and Christoph Scholz. Audioprotopnet: An interpretable deep learning model for bird sound classification. *Ecological Informatics*, 87:103081, 2025. 7, 16

Maximilian HK Hesselbarth, Jakub Nowosad, Alida de Flamingh, Craig E Simpkins, Martin Jung, Gemma Gerber, and Martí Bosch. Computational methods in landscape ecology. *Current Landscape Ecology Reports*, 10:1 – 18, 2025. 3

Sho Hibino, Chifumi Suzuki, and Takanori Nishino. Classification of singing insect sounds with convolutional neural network. *Acoustical Science and Technology*, 42:354 – 356, 2021. 66

Alam Ahmad Hidayat, Tjeng Wawan Cenggoro, and Bens Pardamean. Convolutional neural networks for scops owl sound classification. *Procedia Computer Science*, 179:81 – 87, 2021. 16, 17

Frank Hilker and Frank Westerhoff. Preventing extinction and outbreaks in chaotic populations. *The American naturalist*, 170:232 – 41, 09 2007. doi: 10.1086/518949. 87, 88, 103

Gesa Hoffmann, Aayushi Shukla, Silvia López-González, and Anders Hafrén. Cauliflower mosaic virus disease spectrum uncovers novel susceptibility factor nced9 in arabidopsis thaliana. *Journal of Experimental Botany*, 74:4751 – 4764, 2023. 106

Jim Hone and Charles J Krebs. Causality and wildlife management. *The Journal of Wildlife Management*, page e22412, 2023. 1

Shengbing Hong, Wei Zhan, Tianyu Dong, Jinhui She, Chao Min, Huazi Huang, and Yong Sun. A recognition method of bactrocera minax (diptera: Tephritidae) grooming behavior via a multi-object tracking and spatio-temporal feature detection model. *Journal of Insect Behavior*, 35:67 – 81, 2022. 27

Toke T Høye, Johanna Ärje, Kim Bjerge, Oskar LP Hansen, Alexandros Iosifidis, Florian Leese, Hjalte MR Mann, Kristian Meissner, Claus Melvad, and Jenni Raitoharju. Deep learning and computer vision will transform entomology. *Proceedings of the National Academy of Sciences*, 118, 2021a. 26

Toke T Høye, Johanna Ärje, Kim Bjerge, Oskar LP Hansen, Alexandros Iosifidis, Florian Leese, Hjalte MR Mann, Kristian Meissner, Claus Melvad, and Jenni Raitoharju. Deep learning and computer vision will transform entomology. *Proceedings of the National Academy of Sciences*, 118, 2021b. 3, 4, 66

Fanliang Hu, Jian Shen, and Pandi Vijayakumar. Side-channel attacks based on multi-loss regularized denoising autoencoder. *IEEE Transactions on Information Forensics and Security*, 2023. 74

Yuqing Hu, Sara Si-Moussi, and Wilfried Thuiller. Introduction to deep learning methods for multi-species predictions. *Methods in Ecology and Evolution*, 16: 228 – 246, 2025. 3

Jian Huang and HongFei Hao. Effects of climate change and crop planting structure on the abundance of cotton bollworm, helicoverpa armigera (hübner)(lepidoptera: Noctuidae). *Ecology and evolution*, 10:1324 – 1338, 2020. 8

Jian Huang and Jing Li. Effects of climate change on overwintering pupae of the cotton bollworm, helicoverpa armigera (hübner)(lepidoptera: Noctuidae). *International Journal of Biometeorology*, 59:863 – 876, 2015. 8

Norden E Huang, Zheng Shen, Steven R Long, Manli C Wu, Hsing H Shih, Quanan Zheng, Nai-Chyuan Yen, Chi Chao Tung, and Henry H Liu. The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London. Series A: mathematical, physical and engineering sciences*, 454:903 – 995, 1998. 88

Yo-Ping Huang and Haobijam Basanta. Recognition of endemic bird species using deep learning models. *IEEE Access*, 9:102975 – 102984, 2021. 16

Hiep Xuan Huynh, Duy Bao Lam, Tu Van Ho, Diem Thi Le, and Ly Minh Le. Cdnn model for insect classification based on deep neural network approach. In *Context-Aware Systems and Applications, and Nature of Computation and Communication*, pages 127 – 142. Springer, 2019. 26

Agnes Incze, Henrietta-Bernadett Jancso, Zoltan Szilagyi, Attila Farkas, and Csaba Sulyok. Bird sound recognition using a convolutional neural network. In *Proceedings of the 2018 IEEE 16th International Symposium on Intelligent Systems and Informatics (SISY)*, pages 000295–000300, 09 2018. doi: 10.1109/SISY.2018.8524677. 17

Haris Iqbal. Harisiqbal88/plotneuralnet v1.0.0, December 2018. URL https://doi.org/10.5281/zenodo.2526396. Accessed: 2025-01-08. xvi, xvii, 44, 46, 53

Olaf Jahn, Todor D. Ganchev, Marinez I. Marques, and Karl-L. Schuchmann. Automated sound recognition provides insights into the behavioral ecology of a tropical bird. *PLOS ONE*, 12:1 – 29, 01 2017. doi: 10.1371/journal.pone.0169041. URL https://doi.org/10.1371/journal.pone.0169041. 16

Waqar Jaleel, Lihua Lu, and Yurong He. Biology, taxonomy, and ipm strategies of bactrocera tau walker and complex species (diptera; tephritidae) in asia: a comprehensive review. *Environmental Science and Pollution Research*, 25:19346 – 19361, 2018. 24

Jueun Jeong, Hanseok Jeong, and Han-Joon Kim. An autoencoder-based numerical training data augmentation technique. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 5944 – 5951. IEEE, 2022. 73

Yao Jiang, Zhou Wang, Zhongrui Zhang, Xiaogang Ding, Shaowei Jiang, and Jianguo Huang. Enhancing forest insect outbreak detection by integrating tree-ring and climate variables. *Journal of Forestry Research*, 35, 2024. 107

Nadia Lis Jiménez, Ignacio Raúl Fosco, Gustavo César Nassar, Andrés Fernando Sánchez-Restrepo, Matías Santiago Danna, and Luis Alberto Calcaterra. Economic injury level and economic threshold as required by forest stewardship council for management of leaf-cutting ants in forest plantations. *Agricultural and Forest Entomology*, 23:87 – 96, 2021. 8

Stefan Kahl, Connor M Wood, Maximilian Eibl, and Holger Klinck. Birdnet: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, 61:101236, 2021. 7, 16, 17

Kevin Karbstein, Lara Kösters, Ladislav Hodač, Martin Hofmann, Elvira Hörandl, Salvatore Tomasello, Natascha D Wagner, Brent C Emerson, Dirk C Albach, Stefan Scheu, et al. Species delimitation 4.0: integrative taxonomy meets artificial intelligence. *Trends in Ecology & Evolution*, 39:771 – 784, 2024. 66

Elna Karmawati, Paramita Maris, Rismayani Rismayani, Rohimatun Rohimatun, Gusti Indriati, Dwi Adi Sunarto, Sujak Sujak, Samsudin Samsudin, Iwa Mara Trisawa, Molide Rizal, Siswanto Siswanto, Tri Lestari Mardiningsih, I Gusti Agung Ayu Indrayani, Nurindah Nurindah, Agus Kardinan, and Deciyanto Soetopo. Challenges and constraints in implementing integrated pest management for pepper stem borer (lophobaris piperis marshall) among indonesian smallholder farmers: a critical review. *Journal of Integrated Pest Management*, 16:6, 03 2025. ISSN 2155-7470. doi: 10.1093/jipm/pmaf005. URL https://doi.org/10.1093/jipm/pmaf005. 7

Thenmozhi Kasinathan and Srinivasulu Reddy Uyyala. Machine learning ensemble with image processing for pest identification and classification in field crops. *Neural Computing and Applications*, 33:7491 – 7504, 2021. 25, 26

Thenmozhi Kasinathan, Dakshayani Singaraju, and Srinivasulu Reddy Uyyala. Insect classification and detection in field crops using modern machine learning techniques. *Information Processing in Agriculture*, 8:446 – 457, 2021. 25, 26

Manjit Kaur, Iman Ardekani, Hamid Sharifzadeh, and Soheil Varastehpour. A CNN-based identification of honeybees' infection using augmentation. In *2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, pages 1 – 6. IEEE, 2022. 26

Navpreet Kaur and Amar Singh. Vgg16-pca-pb3c: A hybrid pb3c and deep neural network based approach for leukemia detection. *International Journal of Information Technology*, pages 1 – 11, 2024. 69

Arik Kershenbaum, Çağlar Akçay, Lakshmi Babu-Saheer, Alex Barnhill, Paul Best, Jules Cauzinille, Dena Clink, Angela Dassow, Emmanuel Dufourq, Jonathan Growcott, et al. Automatic detection for bioacoustic research: a practical guide from and for biologists and computer scientists. *Biological Reviews*, 100:620 – 646, 2025. 6

Siti Khairunniza-Bejo, Mohd Firdaus Ibrahim, Marsyita Hanafi, Mahirah Jahari, Fathinul Syahir Ahmad Saad, and Mohammad Aufa Mhd Bookeri. Automatic

paddy planthopper detection and counting using faster r-cnn. *Agriculture*, 14: 1567, 2024. 5

Azal Ahmad Khan, Omkar Chaudhari, and Rohitash Chandra. A review of ensemble learning and data augmentation models for class imbalanced problems: combination, implementation and evaluation. *Expert Systems with Applications*, 244, 2024. 67, 71, 80

M Khedmati, F Seifi, and MJ Azizi. Time series forecasting of bitcoin price based on autoregressive integrated moving average and machine learning approaches. *International Journal of Engineering*, 33:1293 – 1303, 2020. 107

Shivam Khurana. Image enhancement utilizing generative adversarial networks for improving the accuracy of pest insect classification. In *2023 International Conference on Data Science and Network Security (ICDSNS)*, pages 1 – 7. IEEE, 2023. 5, 6

Donghoh Kim, Kyungmee O Kim, and Hee-Seok Oh. Extending the scope of empirical mode decomposition by smoothing. *EURASIP Journal on Advances in Signal Processing*, 2012:1 – 17, 2012. 88

Denis O Kiobia, Canicius J Mwitta, Peter C Ngimbwa, Jason M Schmidt, Guoyu Lu, and Glen C Rains. Machine-learning approach facilitates prediction of whitefly spatiotemporal dynamics in a plant canopy. *Journal of Economic Entomology*, 118:732–745, 02 2025. ISSN 0022-0493. doi: 10.1093/jee/toaf035. URL https://doi.org/10.1093/jee/toaf035. 8

Shigeki Kishi, Jianqiang Sun, Akira Kawaguchi, Sunao Ochi, Megumi Yoshida, and Takehiko Yamanaka. Characteristic features of statistical models and machine learning methods derived from pest and disease monitoring datasets. *Royal Society Open Science*, 10, 2023a. 107

Shigeki Kishi, Jianqiang Sun, Akira Kawaguchi, Sunao Ochi, Megumi Yoshida, and Takehiko Yamanaka. Characteristic features of statistical models and machine learning methods derived from pest and disease monitoring datasets. *Royal Society Open Science*, 10:230079, 2023b. 8

Morris Klasen, Dirk Ahrens, Jonas Eberle, and Volker Steinhage. Image-based automated species identification: can virtual data augmentation overcome problems of insufficient sampling? *Systematic Biology*, 71:320 – 333, 2022. 67

Chih-Yuan Koh, Jaw-Yuan Chang, Chiang-Lin Tai, Da-Yo Huang, Han-Hsing Hsieh, and Yi-Wen Liu. Bird sound classification using convolutional neural networks. In *CLEF (Working Notes)*, 2019. 16

Georgia Koppe, Andreas Meyer-Lindenberg, and Daniel Durstewitz. Deep learning for small and big data in psychiatry. *Neuropsychopharmacology*, 46:176 – 190, 2021. 66

Diamantis Koutsandreas, Evangelos Spiliotis, Fotios Petropoulos, and Vassilios Assimakopoulos. On the selection of forecasting accuracy measures. *Journal of the Operational Research Society*, 73:937 – 954, 2022. 128

Waghmare Kranti, Ghayal Nivedita, and Mahesh Shindikar. Understanding the plant aphid interaction: A review. *European Journal of Biology and Biotechnology*, 2:1 – 6, 2021. 88

Paul R Krausman and James W Cain. *Wildlife management and conservation: contemporary principles and practices*. JHU Press, 2022. 1, 2

Ajay Kumar, Debolina Ghosh, Vivek Kumar, and Jagannath Singh. Deployment of bird-vocal recognition system using deep automatics artificial intelligence. In *International Conference on Mathematical Modeling, Computational Intelligence Techniques and Renewable Energy*, pages 149 – 160. Springer, 2025a. 6

Nitish Kumar, Chirag Sharma, and Tanya Nagpal. Advancing agricultural sustainability: Machine learning-driven insights for optimized crop yield forecasting. *2025 International Conference on Automation and Computation (AUTOCOM)*, 2025b. 2

P Lavanya Kumari, I Paramasiva, U Vineetha, A Veeraiah, Sk Shameem, PN Harathi, ADVSLP Anand Kumar, M Siva Rama Krishna, N Sambasiva Rao, P Udayababu, et al. Climate-driven forecasting of brown planthopper in

rice fields using hybrid machine learning and time series models. *International Journal of Environment and Climate Change*, 15:494 – 514, 2025. 8

Hiroaki Kuzuhara, Hironori Takimoto, Yasuhiro Sato, and Akihiro Kanagawa. Insect pest detection and identification method based on deep learning for realizing a pest control system. In *2020 59th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE)*, pages 709 – 714. IEEE, 2020. 26

Victoria Lantschner, Brian Aukema, and Juan Corley. Droughts drive outbreak dynamics of an invasive forest insect on an exotic host. *Forest Ecology and Management*, 433:762 – 770, 02 2019. doi: 10.1016/j.foreco.2018.11.044. 85, 86

Sedthapong Laojun, Tanasak Changbunjong, Suchada Sumruayphol, and Tanawat Chaiphongpachara. Outline-based geometric morphometrics: Wing cell differences for mosquito vector classification in the tanaosri mountain range, thailand. *Acta Tropica*, 250, 2024. 66

Natalia Larios, Hongli Deng, Wei Zhang, Matt Sarpola, Jenny Yuen, Robert Paasch, Andrew Moldenke, David A Lytle, Salvador Ruiz Correa, Eric N Mortensen, et al. Automated insect identification through concatenated histograms of local appearance features: feature vector generation and region detection for deformable objects. *Machine Vision and Applications*, 19:105 – 123, 2008. 25

Mario Lasseck. Audio-based bird species identification with deep convolutional neural networks. *CLEF (working notes)*, 2125, 2018a. 16

Mario Lasseck. Acoustic bird detection with deep convolutional neural networks. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, pages 143 – 147, 2018b. 16

Douglas Lawton, Anders S. Huseth, George G. Kennedy, Amy C. Morey, William D. Hutchison, Dominic D. Reisig, Seth J. Dorman, DeShae Dillard, Robert C. Venette, Russell L. Groves, John J. Adamczyk, Izailda Barbosa Dos Santos, Tracey Baute, Sebe Brown, Eric Burkness, Ashley Dean, Galen P. Dively, Hélène B. Doughty, Shelby J. Fleischer, Jessica Green, Jeremy K. Greene,

Krista Hamilton, Erin Hodgson, Thomas Hunt, David Kerns, Billy Rogers Leonard, Sean Malone, Fred Musser, David Owens, John C. Palumbo, Silvana Paula-Moraes, Julie A. Peterson, Ricardo Ramirez, Silvia I. Rondon, Tracy L. Schilder, Abby Seaman, Lori Spears, Scott D. Stewart, Sally Taylor, Tyler Towles, Celeste Welty, Joanne Whalen, Robert Wright, and Marion Zuefle. Pest population dynamics are related to a continental overwintering gradient. *Proceedings of the National Academy of Sciences*, 119:e2203230119, 2022. doi: 10.1073/pnas.2203230119. URL https://www.pnas.org/doi/abs/10.1073/pnas.2203230119. 8

Neel P Le Penru, Becky E Heath, Jamie Dunning, Lorenzo Picinali, Robert M Ewers, and Sarab S Sethi. Towards using virtual acoustics for evaluating spatial ecoacoustic monitoring technologies. *Methods in Ecology and Evolution*, 16:108 – 125, 2025. 7

Simon R Leather. Influential entomology: a short review of the scientific, societal, economic and educational services provided by entomology. *Ecological Entomology*, 40:36 – 44, 2015. 3, 4

Simon R Leather. "Ecological Armageddon"-more evidence for the drastic decline in insect numbers. *Annals of Applied Biology*, 172:1 – 3, 2017. 2

Jack LeBien, Ming Zhong, Marconi Campos-Cerqueira, Julian P Velev, Rahul Dodhia, Juan Lavista Ferres, and T Mitchell Aide. A pipeline for identification of bird and frog species in tropical soundscape recordings using a convolutional neural network. *Ecological Informatics*, 59:101113, 2020. 16

Yann LeCun, Koray Kavukcuoglu, and Clément Farabet. Convolutional networks and applications in vision. In *Proceedings of 2010 IEEE international symposium on circuits and systems*, pages 253 – 256. IEEE, 2010. 3

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521: 436 – 444, 2015. 3

Yann LeCun et al. Generalization and network design strategies. *Connectionism in perspective*, 19, 1989. 26

Matheus Macedo Leonardo, Tiago J Carvalho, Edmar Rezende, Roberto Zucchi, and Fabio Augusto Faria. Deep feature-based classifiers for fruit fly identification (diptera: Tephritidae). In *2018 31st SIBGRAPI conference on graphics, patterns and images (SIBGRAPI)*, pages 41 – 47. IEEE, 2018a. 25

Matheus Macedo Leonardo, Tiago J Carvalho, Edmar Rezende, Roberto Zucchi, and Fabio Augusto Faria. Deep feature-based classifiers for fruit fly identification (diptera: Tephritidae). In *2018 31st SIBGRAPI conference on graphics, patterns and images (SIBGRAPI)*, pages 41 – 47. IEEE, 2018b. 67, 81

Miguel Lerma and Mirtha Lucas. Grad-cam++ is equivalent to grad-cam with positive gradients. In *Proceedings of the Irish Machine Vision and Image Processing Conference*, pages 113 – 120, 2022. 51

Daniel J Leybourne, Nasamu Musa, and Po Yang. Can artificial intelligence be integrated into pest monitoring schemes to help achieve sustainable agriculture? an entomological, management and computational perspective. *Agricultural and Forest Entomology*, 27:8 – 17, 2025. 3, 4

Guozhi Li, Zhengbo Liu, Zelin Feng, Jun Lyu, Bin Li, Guo Chen, and Qing Yao. Research on a machine vision-based electro-killing pheromone-baited intelligent agricultural pest monitoring method. *Frontiers in Plant Science*, 16:1521594, 2025. 4, 5

Yong Li and Shihua Gong. Dynamic ant colony optimisation for tsp. *The International Journal of Advanced Manufacturing Technology*, 22:528 – 533, 2003. 71

Richard L Lindroth, Mark R Zierden, Clay J Morrow, and Patricia C Fernandez. Forest defoliation by an invasive outbreak insect: Catastrophic consequences for a charismatic mega moth. *Ecology and Evolution*, 14, 2024. 106

Ming Liu, Qiyu Sun, Dustin E Brewer, Thomas M Gehring, and Jesse Eickholt. An ornithologist's guide for including machine learning in a workflow to identify a secretive focal species from recorded audio. *Remote Sensing*, 14:3816, 2022. 7

Qing Liu and Donald A. Pierce. A note on Gauss—Hermite quadrature. *Biometrika*, 81:624 – 629, 09 1994. ISSN 0006-3444. doi: 10.1093/biomet/81.3.624. URL https://doi.org/10.1093/biomet/81.3.624. 94

Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma. A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 869 – 884. Springer, 2016. 3

Yang Liu, Guoping Yang, Shaojie Qiao, Meiqi Liu, Lulu Qu, Nan Han, Tao Wu, Guan Yuan, and Yuzhong Peng. Imbalanced data classification: Using transfer learning and active sampling. *Engineering Applications of Artificial Intelligence*, 117, 2023. 67

Vincent Lostanlen, Justin Salamon, Andrew Farnsworth, Steve Kelling, and Juan Pablo Bello. Robust sound event detection in bioacoustic sensor networks. *PLOS ONE*, 14:1–31, 10 2019. doi: 10.1371/journal.pone.0214168. URL https://doi.org/10.1371/journal.pone.0214168. 7

Vadim Lozin. The hamiltonian cycle problem and monotone classes. In *International Workshop on Combinatorial Algorithms*, pages 460 – 471. Springer, 2024. 70

Sarah H Luke, Helen E Roy, Chris D Thomas, Luke AN Tilley, Simon Ward, Allan Watt, Manuela Carnaghi, Coline C Jaworski, Maximillian PTG Tercel, Charlie Woodrow, et al. Grand challenges in entomology: Priorities for action in the coming decades. *Insect conservation and diversity*, 16:173 – 189, 2023. 3, 4

Erick J Lundgren, Arian D Wallach, Jens-Christian Svenning, Martin A Schlaepfer, Astrid LA Andersson, and Daniel Ramp. Preventing extinction in an age of species migration and planetary change. *Conservation Biology*, page e14270, 2024. 2

Lynch. Spruce aphid, elatobium abietinum (walker): Life history and damage to engelmann spruce in the pinaleno mountains, arizona. *The Last Refuge of the Mt. Graham Red Squirrel: Ecology of Endangerment*, 01 2009. 86

Lynch. Socioecological impacts of multiple forest insect outbreaks in the pinaleño spruce–fir forest, arizona. *Journal of Forestry*, 117, 10 2018. doi: 10.1093/jofore/fvy039. 86

Renato R Maaliw, Melvin A Ballera, Zoren P Mabunga, Aubee T Mahusay, Dhenalyn A Dejelo, and Mariebeth P Seño. An ensemble machine learning approach for time series forecasting of covid-19 cases. In *2021 IEEE 12th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pages 0633 – 0640. IEEE, 2021. 107

Sarina Macfadyen and Darren Kriticos. Modelling the geographical range of a species with variable life-history. *Public Library of Science One*, 7, 07 2012. doi: 10.1371/journal.pone.0040313. 97

Laurence Madden and M Wheelis. The threat of plant pathogens as weapons against U.S. crops. *Annual review of phytopathology*, 41:155 – 76, 02 2003. doi: 10.1146/annurev.phyto.41.121902.102839. 84

Akila Maithripala, Samantha Mathara Arachchi, Kasun Karunanayaka, Ravindu Perera, and Pandula Pallewatta. A review of automated bird sound recognition and analysis in the new AI era. In *2024 8th SLAAI International Conference on Artificial Intelligence (SLAAI-ICAI)*, pages 1 – 6. IEEE, 2024. 6

Aldo Malavasi and Roberto Antônio Zucchi, editors. *Moscas-das-frutas de importância econômica no Brasil: conhecimento básico e aplicado*. Holos, Ribeirão Preto, 2000. ISBN 8586-6991-36. 67

Nadia Mansouri and Zied Lachiri. Laughter synthesis: A comparison between variational autoencoder and autoencoder. In *2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, pages 1 – 6. IEEE, 2020. 73

Maxime Martineau, Donatello Conte, Romain Raveaux, Ingrid Arnault, Damien Munier, and Gilles Venturini. A survey on image-based insect classification. *Pattern Recognition*, 65:273 – 284, 2017. 26

María Martínez-Jauregui, Miguel Delibes-Mateos, Beatriz Arroyo, Jenny Anne Glikman, and Mario Soliño. Beyond rural vs urban differences: A close match in European preferences in some basic wildlife management and conservation principles. *Journal of Environmental Management*, 331:117236, 2023. 1

Valter AM Martins, Lucas C Freitas, Marilton S de Aguiar, Lisane B de Brisolara, and Paulo R Ferreira. Deep learning applied to the identification of fruit fly in intelligent traps. In *2019 IX Brazilian Symposium on Computing Systems Engineering (SBESC)*, pages 1 – 8. IEEE, 2019a. 25

Valter AM Martins, Lucas C Freitas, Marilton S de Aguiar, Lisane B de Brisolara, and Paulo R Ferreira. Deep learning applied to the identification of fruit fly in intelligent traps. In *2019 IX Brazilian symposium on computing systems engineering (SBESC)*, pages 1 – 8. IEEE, 2019b. 67

Sheldon Mascarenhas and Mukul Agarwal. A comparison between vgg16, vgg19 and resnet50 architecture frameworks for image classification. In *2021 International conference on disruptive technologies for multi-disciplinary research and applications (CENTCON)*, volume 1, pages 96 – 99. IEEE, 2021. 3

Ricardo P Masini, Marcelo C Medeiros, and Eduardo F Mendes. Machine learning advances for time series forecasting. *Journal of economic surveys*, 37:76 – 111, 2023. 107, 115

Rubén Mateos Fernández, Marko Petek, Iryna Gerasymenko, Mojca Juteršek, Špela Baebler, Kalyani Kallam, Elena Moreno Giménez, Janine Gondolf, Alfred Nordmann, Kristina Gruden, et al. Insect pest management in the age of synthetic biology. *Plant Biotechnology Journal*, 20:25 – 36, 2022. 106, 107

Aurore Mathys, Yann Pollet, Adrien Gressin, Xavier Muth, Jonathan Brecko, Wouter Dekoninck, Didier Vandenspiegel, Sébastien Jodogne, and Patrick Semal. Sphaeroptica: a tool for pseudo-3d visualization and 3d measurements on arthropods. *Plos one*, 19:e0311887, 2024. 5

K Sherin ME, AR Darshika Kelin ME, Surendar Senthilvelan, et al. Birds species identification using deep learning model. In *2024 International Conference on*

Advances in Data Engineering and Intelligent Computing Systems (ADICS), pages 1 – 7. IEEE, 2024. 7

Carlos Medina-Ramos, Henrique De Aguiar-Minami, Daniel Carbonel-Olazabal, Jhon Zelada-Rodriguez, Michael Vera-Panez, and Alonso Tenorio-Trigoso. Object detection algorithms identifying ceratitis capitata fruit fly. In *2024 IEEE Biennial Congress of Argentina (ARGENCON)*, pages 1 – 8. IEEE, 2024. 67

Rodrigo Mello and Moacir Ponti. *Machine Learning: A Practical Approach on the Statistical Learning Theory*. Springer, 01 2018. ISBN 978-3-319-94988-8. doi: 10.1007/978-3-319-94989-5. 25, 37, 108

Zhongqi Miao, Ziwei Liu, Kaitlyn M Gaynor, Meredith S Palmer, Stella X Yu, and Wayne M Getz. Iterative human and automated identification of wildlife images. *Nature Machine Intelligence*, 3:885 – 895, 2021. 5, 6

Paul Mitchell and David Onstad. *Valuing Pest Susceptibility to Control*, pages 17–38. Elsevier Science, 01 2014. ISBN 9780123738585. doi: 10.1016/B978-012373858-5.50004-6. 2

Paul Mitchell, Michael Gray, and Kevin Steffey. A composed-error model for estimating pest-damage functions and the impact of the western corn rootworm soybean variant in illinois. *American Journal of Agricultural Economics*, 86:332 –344, 02 2004. doi: 10.1111/j.0092-5853.2004.00582.x. 87

Mahsa Mohaghegh, Khaula Alizai, Minh Hoang, Kapil Patel, and June Lee. Machine learning for conservation: Evaluating deep learning and feature extraction in bird species classification in new zealand. In *2023 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, pages 1–6, 2023. doi: 10.1109/CSDE59766.2023.10487683. 7

B Chandra Mohan and R Baskaran. A survey: Ant colony optimization based recent research and implementation on several engineering domain. *Expert Systems with Applications*, 39:4618 – 4627, 2012. 71

Ingrid Molina-Mora, Viviana Ruíz-Gutierrez, Álvaro Vega-Hidalgo, and Luis Sandoval. The utility of passive acoustic monitoring for using birds as indicators

of sustainable agricultural management practices. *Frontiers in Bird Science*, 3: 1386759, 2024. 6

Miguel Molina-Rotger, Alejandro Morán, Miguel Angel Miranda, and Bartomeu Alorda-Ladaria. Remote fruit fly detection using computer vision and machine learning-based electronic trap. *Frontiers in Plant Science*, 14, 2023. 67

Abdullah Moonis and Ajeet Singh. Optimized insect classification on farms using tuned convolutional neural networks. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1 – 5. IEEE, 2024. 66

Katharine M Mullen. Continuous global optimization in r. *Journal of Statistical Software*, 60:1 – 45, 2014. 94

Jörg Müller, Heinz Bußler, Martin Goßner, Thomas Rettelbach, and Peter Duelli. The European spruce bark beetle ips typographus in a national park: from pest to keystone species. *Biodiversity and Conservation*, 17:2979 – 3001, 2008. 106

Darrell RJ Mullowney and Krista D Baker. Multi-indicator precautionary approach frameworks for crustacean fisheries. *Canadian Journal of Fisheries and Aquatic Sciences*, 80:1207 – 1220, 2023. 2

Alhassan Mumuni, Fuseini Mumuni, and Nana Kobina Gerrar. A survey of synthetic data augmentation methods in machine vision. *Machine Intelligence Research*, pages 1 – 39, 2024. 66

K.S.S. Nair. Pest outbreaks in tropical forest plantations: is there a greater risk for exotic tree species? *Center for International Forestry Research*, pages 1 – 82, 01 2001. doi: 10.17528/cifor/000984. 85, 86

K.S.S. Nair. Tropical forest insect pests. ecology, impact, and management. *Tropical Forest Insect Pests: Ecology, Impact, and Management*, pages 1 – 404, 01 2007. doi: 10.1017/CBO9780511542695. 85, 86

Loris Nanni, Nicola Maritan, Daniel Fusaro, Sheryl Brahnam, Francesco Boscolo Meneguolo, and Maria Sgaravatto. Insect identification by combining different

neural networks. *Expert Systems with Applications*, 273, 2025. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2025.126935. URL https://www.sciencedirect.com/science/article/pii/S0957417425005573. 5

Vahid Nasir and Farrokh Sassani. A review on deep learning in machining and tool monitoring: Methods, opportunities, and challenges. *The International Journal of Advanced Manufacturing Technology*, 115:2683 – 2709, 2021. 3

AP Naufal, C Kanjanaphachoat, A Wijaya, NA Setiawan, and RE Masithoh. Insects identification with convolutional neural network technique in the sweet corn field. In *IOP Conference Series: Earth and Environmental Science*, volume 653. IOP Publishing, 2021. 25

Sarah Nawoya, Frank Ssemakula, Roseline Akol, Quentin Geissmann, Henrik Karstoft, Kim Bjerge, Cosmas Mwikirize, Andrew Katumba, and Grum Gebreyesus. Computer vision and deep learning in insects for food and feed production: A review. *Computers and Electronics in Agriculture*, 216, 2024. 66

Jose Negron, Barbara Bentz, Christopher Fettig, Nancy Gillette, E. Hansen, Jane Hayes, Rick Kelsey, John Lundquist, Ann Lynch, Robert Progar, and Steven Seybold. Us forest service bark beetle research in the western united states: Looking toward the future. *Journal of Forestry*, 106:325 – 331, 09 2008. 86

Nitin, Satinder Bal Gupta, RajKumar Yadav, Fatemeh Bovand, and Pankaj Kumar Tyagi. Developing precision agriculture using data augmentation framework for automatic identification of castor insect pests. *Frontiers in Plant Science*, 14, 2023. 67

Allen L. Norrbom, Roberto A. Zucchi, and Vicente Hernández-Ortiz. Phylogeny of the genera anastrepha and toxotrypana (trypetinae: Toxotrypanini) based on morphology. In Martin Aluja and Allen L. Norrbom, editors, *Fruit Flies (Tephritidae): Phylogeny and Evolution of Behavior*, pages 317 – 360. CRC Press, Boca Raton, 1999. 67

G. W. Barrett Odum. *Fundamentals of Ecology*. BioOne, 02 2005. 85, 86

Mitchell O'Hara-Wild, Rob J Hyndman, Earo Wang, Christoph Bergmeir, Gabriel Caceres, Tim-Gunnar Hensel, Thomas Petzoldt, and Earo Wang. *fable: Forecasting Models for Tidy Time Series*, 2023. URL https://cran.r-project.org/web/packages/fable/fable.pdf. R package version 0.3.3. Accessed: 2025-01-08. 113

Song-Quan Ong and Hamdan Ahmad. An annotated image dataset of medically and forensically important flies for deep learning model training. *Scientific Data*, 9:1 – 7, 2022. xv, xxiv, 10, 25, 31, 34, 35

Song-Quan Ong and Toke Thomas Høye. Trap colour strongly affects the ability of deep learning models to recognize insect species in images of sticky traps. *Pest Management Science*, 81:654 – 666, 2025. 5

Maria Angélica Ono, Elisângela Novais Lopes Ferreira, and Wesley Augusto Conde Godoy. Black wattle insect pests currently in Brazil. *Glo Adv Res J Agric Sci*, 3:409 – 414, 2014. 86

David W. Onstad. Calculation of Economic-injury Levels and Economic Thresholds for Pest Management. *Journal of Economic Entomology*, 80:297 – 303, 04 1987. ISSN 0022-0493. doi: 10.1093/jee/80.2.297. URL https://doi.org/10.1093/jee/80.2.297. 87

Durmus Ozdemir and Musa Selman Kunduraci. Comparison of deep learning techniques for classification of the insects in order level with mobile software application. *IEEE Access*, 10:35675 – 35684, 2022. 25

Roop Pahuja and Avijeet Kumar. Sound-spectrogram based automatic bird species recognition using mlp classifier. *Applied Acoustics*, 180:108077, 2021. 7

Gabriel Palma, Charles Markham, and Moral Rafael. Detecting predation interaction using pretrained CNNs. In *Proceedings of the Irish Machine Vision and Image Processing Conference*, pages 17 – 20, 2020. 51

Gabriel R. Palma, Ana C. M. M. Aquino, Patricia F. Monticelli, Luciano M. Verdade, Charles Markham, and Rafael A. Moral. A machine vision system for avian song classification with CNN's. In Richard Gault, editor, *Proceedings of*

the 24th Irish Machine Vision and Image Processing conference, pages 64 – 71. Irish Pattern Recognition & Classification Society, September 2022. 27, 69

Gabriel R Palma, Wesley AC Godoy, Eduardo Engel, Douglas Lau, Edgar Galvan, Oliver Mason, Charles Markham, and Rafael A Moral. Pattern-based prediction of population outbreaks. *Ecological Informatics*, 77, 2023a. 107, 115

Gabriel R Palma, Conor P Hackett, and Charles Markham. Machine vision applied to entomology. In *Modelling Insect Populations in Agricultural Landscapes*, pages 149 – 184. Springer, 2023b. 66, 70

Daniel Paredes, Jay A Rosenheim, and Daniel S Karp. The causes and consequences of pest population variability in agricultural landscapes. *Ecological Applications*, 32:e2607, 2022. 8

Carlos Parra-López, Saker Ben Abdallah, Guillermo Garcia-Garcia, Abdo Hassoun, Pedro Sánchez-Zamora, Hana Trollman, Sandeep Jagtap, and Carmen Carmona-Torres. Integrating digital technologies in agriculture for climate change adaptation and mitigation: State of the art and future perspectives. *Computers and Electronics in Agriculture*, 226:109412, 2024. 2

Athanasios Passias, Karolos-Alexandros Tsakalos, Nick Rigogiannis, Dionisis Voglitsis, Nick Papanikolaou, Maria Michalopoulou, George Broufas, and Georgios Ch Sirakoulis. Insect pest trap development and dl-based pest detection: A comprehensive review. *IEEE Transactions on AgriFood Electronics*, 2024. 66

Deven Patel and Nirav Bhatt. Improved accuracy of pest detection using augmentation approach with faster r-cnn. In *IOP Conference Series: Materials Science and Engineering*, volume 1042. IOP Publishing, 2021. 25

Judea Pearl et al. Models, reasoning and inference. *Cambridge, UK: Cambridge-UniversityPress*, 19:3, 2000. 109

Martín Pedemonte, Sergio Nesmachnow, and Héctor Cancela. A survey on parallel ant colony optimization. *Applied Soft Computing*, 11:5181 – 5197, 2011. 71

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825 – 2830, 2011. Accessed: 2025-01-08. 75

Yingqiong Peng, Muxin Liao, Yuxia Song, Zhichao Liu, Huojiao He, Hong Deng, and Yinglong Wang. Fb-cnn: Feature fusion-based bilinear cnn for classification of fruit fly image. *IEEE Access*, 8:3987 – 3995, 2019. 25

Silvester Dian Handy Permana, Gusti Saputra, Budi Arifitama, Wahyu Caesarendra, Robbi Rahim, et al. Classification of bird sounds as an early warning method of forest fires using convolutional neural network (cnn) algorithm. *Journal of King Saud University-Computer and Information Sciences*, 2021. 16, 17

P. Perre, F. A. Faria, L. R. Jorge, et al. Toward an automated identification of anastrepha fruit flies in the fraterculus group (diptera, tephritidae). *Neotropical Entomology*, 45:554 – 558, 2016. doi: 10.1007/s13744-016-0403-0. URL https://doi.org/10.1007/s13744-016-0403-0. 81

Thomas M Perring, Ned M Gruenhagen, and Charles A Farrar. Management of plant viral diseases through chemical control of insect vectors. *Annual review of entomology*, 44:457 – 481, 1999. 106

Tinao Petso, Rodrigo S. Jamisola, Dimane Mpoeleng, Emily Bennitt, and Wazha Mmereki. Automatic animal identification from drone camera based on point pattern analysis of herd behaviour. *Ecological Informatics*, 66, 2021. ISSN 1574-9541. doi: https://doi.org/10.1016/j.ecoinf.2021.101485. URL https://www.sciencedirect.com/science/article/pii/S1574954121002764. 2

Bui Hai Phong, Nguyen Thi Hong Thuy, and Pham Hoang Quan. A classification method for insects using data augmentation and deep neural networks. *International Journal of Advanced Research in Computer Science*, 15, 2024. 67

Mutondwa Phophi, Paramu Mafongoya, and Shenelle Lottering. Perceptions of climate change and drivers of insect pest outbreaks in vegetable crops in limpopo province of south africa. *Climate*, 8, 02 2019. doi: 10.3390/cli8020027. 86

Quoc Viet Phung, Iftekhar Ahmad, Daryoush Habibi, and Steven Hinckley. Automated insect detection using acoustic features based on sound generated from insect activities. *Acoustics Australia*, 45:445 – 451, 2017. 66

Francesca Pianosi, Keith Beven, Jim Freer, Jim W Hall, Jonathan Rougier, David B Stephenson, and Thorsten Wagener. Sensitivity analysis of environmental models: A systematic review with practical workflow. *Environmental Modelling & Software*, 79:214 – 232, 2016. 105

Bryan Pijanowski, Luis Villanueva-Rivera, Sarah Dumyahn, Almo Farina, Bernie Krause, Brian Napoletano, Stuart Gage, and Nadia Pieretti. Soundscape ecology: The science of sound in the landscape. *BioScience*, 61, 03 2011. doi: 10.1525/bio.2011.61.3.6. 15

Bryan C Pijanowski, Francisco Rivas Fuenzalida, Subham Banerjee, Rosane Minghim, Samantha L Lima, Ruth Bowers-Sword, Santiago Ruiz Guzman, Josept Revuelta-Acosta, Adebola Esther Adeniji, Sarah E Grimes, et al. Soundscape analytics: A new frontier of knowledge discovery in soundscape data. *Current Landscape Ecology Reports*, 9:88 – 107, 2024. 6

RESHMA PISE and KAILAS PATIL. Imbalanced class learning in vision based classification of vector mosquito species. *Journal of Theoretical and Applied Information Technology*, 102, 2024. 66, 67

Ilyas Potamitis. Unsupervised dictionary extraction of bird vocalisations and new tools on assessing and visualising bird activity. *Ecological Informatics*, 26:6 – 17, 2015. 6

David Pratella, Samira Ait-El-Mkadem Saadi, Sylvie Bannwarth, Véronique Paquis-Fluckinger, and Silvia Bottini. A survey of autoencoder algorithms to pave the diagnosis of rare diseases. *International journal of molecular sciences*, 22, 2021. 71

Nirosha Priyadarshani, Stephen Marsland, and Isabel Castro. Automated birdsong recognition in complex acoustic environments: a review. *Journal of Avian Biology*, 49. 6

YAO Qing, FENG Jin, TANG Jian, Wei-gen XU, Xu-hua ZHU, Bao-jun YANG, LÜ Jun, Yi-ze XIE, YAO Bo, Shu-zhen WU, et al. Development of an automatic monitoring system for rice light-trap pests based on machine vision. *Journal of Integrative Agriculture*, 19:2500 – 2513, 2020. 25

Jawwad A Qureshi, Michael E Rogers, David G Hall, and Philip A Stansly. Incidence of invasive diaphorina citri (hemiptera: Psyllidae) and its introduced parasitoid tamarixia radiata (hymenoptera: Eulophidae) in florida citrus. *Journal of Economic Entomology*, 102:247 – 256, 2009. 24

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2025. URL https://www.R-project.org/. Accessed: 2025-01-08. 12, 113

Rahadian Kristiyanto Rachman, De Rosal Ignatius Moses Setiadi, Ajib Susanto, Kristiawan Nugroho, and Hussain Md Mehedul Islam. Enhanced vision transformer and transfer learning approach to improve rice disease recognition. *Journal of Computing Theories and Applications*, 1:446 – 460, 2024. 66

Francesca Raffini, Giorgio Bertorelle, Roberto Biello, Guido D'Urso, Danilo Russo, and Luciano Bosso. From nucleotides to satellite imagery: Approaches to identify and manage the invasive pathogen xylella fastidiosa and its insect vectors in europe. *Sustainability*, 12, 2020. 66

David W Ragsdale, BP McCornack, RC Venette, Bruce D Potter, Ian V MacRae, Erin W Hodgson, Matthew E O'Neal, Kevin D Johnson, RJ O'neil, CD DiFonzo, et al. Economic threshold for soybean aphid (hemiptera: Aphididae). *Journal of Economic Entomology*, 100:1258 – 1267, 2007. 7

Ali Rajabpour. Evaluation of selected biorational and synthetic chemical insecticides for controlling phenacoccus solenopsis (hemiptera: Pseudococcidae) under field conditions. *Journal of Plant Diseases and Protection*, 132:1 – 7, 2025. 7

Pouria Ramazi, Mélodie Kunegel-Lion, Russell Greiner, and Mark A Lewis. Predicting insect outbreaks using machine learning: A mountain pine beetle case study. *Ecology and evolution*, 11:13014 – 13028, 2021. 85, 107

Marine Randon, Michael Dowd, and Ruth Joy. A real-time data assimilative forecasting system for animal tracking. *Ecology*, 103:e3718, 2022. 2

Allan Rodrigues Rebelo, Joao Marcos Garcia Fagundes, Luciano Antonio Digiampietri, and Helton Hideraldo Bíscaro. Methods for automatic image-based classification of winged insects using computational techniques: A systematic literature review. In *Proceedings of the XVI Brazilian Symposium on Information Systems*, SBSI '20, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450388733. doi: 10.1145/3411564.3411641. URL https://doi.org/10.1145/3411564.3411641. 5

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779 – 788, 2016. 54

James R Reilly and Bret D Elderd. Effects of biological control on long-term population dynamics: identifying unexpected outcomes. *Journal of Applied Ecology*, 51:90 – 101, 2014. 7

Dominic Reisig and Anders Huseth. Establishing an ipm system for tarnished plant bug (hemiptera: Miridae) in north carolina. *Insects*, 16, 2025. ISSN 2075-4450. doi: 10.3390/insects16020164. URL https://www.mdpi.com/2075-4450/16/2/164. 8

A Revathi and N Sasikaladevi. Robust sound-based bird classification using multiple features and random forest classifier. *International Journal of Speech Technology*, pages 1 – 11, 2025. 6

Salim Rezvani and Xizhao Wang. A broad review on class imbalance learning techniques. *Applied Soft Computing*, 143, 2023. 67, 80

K Rimal, KB Shah, and AK Jha. Advanced multi-class deep learning convolution neural network approach for insect pest classification using tensorflow. *International Journal of Environmental Science and Technology*, pages 1 – 14, 2022. 26

Ricardo Araújo Rios and Rodrigo Fernandes de Mello. Improving time series modeling by decomposing and analyzing stochastic and deterministic influences. *Signal Processing*, 93:3001 – 3013, 2013. ISSN 0165-1684. doi: https://doi.org/10.1016/j.sigpro.2013.04.017. URL https://www.sciencedirect.com/science/article/pii/S0165168413001606. 109

Federico Riva, Nick Haddad, Lenore Fahrig, and Cristina Banks-Leite. Principles for area-based biodiversity conservation. *Ecology Letters*, 27, 2024. 65

Bruno Leite Rodrigues, Glaucilene da Silva Costa, Rodrigo Espíndola Godoy, Antonio Marques Pereira Júnior, Wilsandrei Cella, Gabriel Eduardo Melim Ferreira, Jansen Fernandes de Medeiros, and Paloma Helena Fernandes Shimabukuro. Molecular and morphometric study of Brazilian populations of psychodopygus davisi. *Medical and Veterinary Entomology*, 38:83 – 98, 2024. 66

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234 – 241, Cham, 2015a. Springer International Publishing. ISBN 978-3-319-24574-4. 52

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234 – 241. Springer, 2015b. 26

Sheldon M. Ross. *A First Course in Probability*. Prentice Hall, Upper Saddle River, N.J., fifth edition, 1998. 85

Abdelhak Rouabah, Helmut Meiss, Jean Villerd, Françoise Lasserre-Joulin, Véronique Tosser, André Chabert, and Olivier Therond. Predicting the abundances of aphids and their natural enemies in cereal crops: Machine-learning versus linear models. *Biological Control*, 169, 2022. 107

DB Roy, J Alison, TA August, M Bélisle, K Bjerge, JJ Bowden, MJ Bunsen, F Cunha, Q Geissmann, K Goldmann, et al. Towards a standardized frame-

work for AI-assisted, image-based monitoring of nocturnal insects. *Philosophical Transactions of the Royal Society B*, 379, 2024. 5, 6

Zachary J Ruff, Damon B Lesmeister, Cara L Appel, and Christopher M Sullivan. A convolutional neural network and r-shiny app for automated identification and classification of animal sounds. *bioRxiv*, 2020. 16

Zachary J Ruff, Damon B Lesmeister, Cara L Appel, and Christopher M Sullivan. Workflow and convolutional neural network for automated identification of animal sounds. *Ecological Indicators*, 124:107419, 2021. 16, 17

Jakob Runge. Modern causal inference approaches to investigate biodiversity-ecosystem functioning relationships. *Nature Communications*, 14:1917, April 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-37546-1. URL https://www.nature.com/articles/s41467-023-37546-1. 8

Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science advances*, 5, 2019a. 105

Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science advances*, 5, 2019b. 129

Jakob Runge, Andreas Gerhardus, Gherardo Varando, Veronika Eyring, and Gustau Camps-Valls. Causal inference for time series. *Nature Reviews Earth & Environment*, 4:487 – 505, 2023. 129

Erik Saberski, Tom Lorimer, Delia Carpenter, Ethan Deyle, Ewa Merz, Joseph Park, Gerald M. Pao, and George Sugihara. The impact of data resolution on dynamic causal inference in multiscale ecological networks. *Communications Biology*, 7, Nov 2024. doi: 10.1038/s42003-024-07054-z. 8

Daisy Salifu, Eric Ali Ibrahim, and Henri EZ Tonnang. Leveraging machine learning tools and algorithms for analysis of fruit fly morphometrics. *Scientific reports*, 12:1 – 11, 2022. 26

E Sankarganesh. Insect biodiversity: The teeming millions-a review. *Bull Environ Pharmacol Life Sci*, 6:101 – 5, 2017. 65

Mangalam Sankupellay and Dmitry Konovalov. Bird call recognition using deep convolutional neural network, resnet-50. In *Proc. Acoustics*, volume 7, pages 1 – 8, 2018. 16

Sabrina Santos, Alexandre Specht, Eduardo Carneiro, Silvana Paula-Moraes, and Mirna Casagrande. Interseasonal variation of chrysodeixis includens (walker, [1858]) (lepidoptera: Noctuidae) populations in the Brazilian savanna. *Revista Brasileira de Entomologia*, 61, 07 2017. doi: 10.1016/j.rbe.2017.06.006. 85, 86

Troy Day Sarah P. Otto. *A Biologist's Guide to Mathematical Modeling in Ecology and Evolution*. Princeton University Press, USA, 1rd edition, 2007. ISBN 9780691123448. 85, 94

Saty Saran, Shivanand S. Hiremath, Akhilesh Kumar, Ashoka P, Harpal Singh, Shaon Chakraborty, Vivek Kashyap, Awanindra Kumar Tiwari, and Shivam Kumar Pandey. Remote sensing and automated monitoring systems for insect pest detection and surveillance. *UTTAR PRADESH JOURNAL OF ZOOLOGY*, 46:155 – 171, January 2025. ISSN 0256-971X, 0256-971X. doi: 10.56557/upjoz/2025/v46i24771. URL https://mbimph.com/index.php/UPJOZ/article/view/4771. 4, 5

Iqbal H Sarker. Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. *SN computer science*, 2:1 – 20, 2021. 3

Andy Sarroff and Michael A Casey. Musical audio synthesis using autoencoding neural nets. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR2014)*. International Society for Music Information Retrieval, 2014. 71

Juan M Scavuzzo, Francisco Trucco, Manuel Espinosa, Carolina B Tauro, Marcelo Abril, Carlos M Scavuzzo, and Alejandro C Frery. Modeling dengue vector population using remotely sensed data and machine learning. *Acta tropica*, 185: 167 – 175, 2018. 107

Stefan Schneider, Graham W Taylor, Stefan C Kremer, and John M Fryxell. Getting the bugs out of AI: Advancing ecological research on arthropods through computer vision. *Ecology Letters*, 26:1247 – 1258, 2023. 4, 5

Jaume Segura-Garcia, Sean Sturley, Miguel Arevalillo-Herraez, Jose M Alcaraz-Calero, Santiago Felici-Castell, and Enrique A Navarro-Camba. 5G AI-IoT system for bird species monitoring and song classification. *Sensors*, 24:3687, 2024. 7

Arja Selin, Jari Turunen, and Juha T Tanttu. Wavelets in recognition of bird sounds. *EURASIP Journal on Advances in Signal Processing*, 2007:1 – 9, 2006. 16

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618 – 626, 2017. 51

JoonOh Seo, SangUk Han, SangHyun Lee, and Hyoungkwan Kim. Computer vision techniques for construction safety and health monitoring. *Advanced Engineering Informatics*, 29:239 – 251, 2015. 3

Wahyudi Setiawan and Riries Rulaningtyas. Visual explanation of maize leaf diaseases classification using squeezenet and gradient-weighted class activation map. In *AIP Conference Proceedings*, volume 2679. AIP Publishing LLC, 2023. 51

Ajay Sharma, Rajneesh Kumar Patel, Pranshu Pranjal, Bhupendra Panchal, and Siddharth Singh Chouhan. Computer vision-based smart monitoring and control system for crop. In *Applications of computer vision and drone technology in agriculture 4.0*, pages 65 – 82. Springer, 2024. 3

Hari C Sharma and Mukesh K Dhillon. Climate change effects on arthropod diversity and its implications for pest management and sustainable crop production. *Agroclimatology: Linking agriculture to climate*, 60:595 – 619, 2018. 86

Sandhya Sharma, Kazuhiko Sato, and Bishnu Prasad Gautam. A methodological literature review of acoustic wildlife monitoring using artificial intelligence tools and techniques. *Sustainability*, 15:7128, 2023. 7

Sarowar Morshed Shawon, Falguny Barua Ema, Asura Khanom Mahi, and Md. Mohsin Sarker Raihan. Crop yield prediction: Robust machine learning approaches for precision agriculture. *2023 26th International Conference on Computer and Information Technology (ICCIT)*, 2023. 2

Y Shen, MZ Hossain, S Rahman, and KA Ahmed. Systematics of tephritid fruit flies: A machine learning based pest identification system. proceedings 2021, 68, 0, 2021. 25

Yefeng Shen, Md Zakir Hossain, Khandaker Asif Ahmed, and Shafin Rahman. An open set model for pest identification. *Computational Biology and Chemistry*, 108, 2024. 67

Nishith S Shetty, Naman N Karanth, and Vinay Hegde. Enhancing crop productivity: Integrated solutions for crop yield, crop recommendation, and crop disease management. *2023 7th International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)*, 2023. 2

Zhichao Shi, Hao Dang, Zhicai Liu, and Xiaoguang Zhou. Detection and identification of stored-grain insects using deep learning: A more effective neural network. *IEEE Access*, 8:163703 – 163714, 2020. 26

M. E. Shimbori1, B. R. Querino, A. V. Costa, and Zucchi A. R. Taxonomy and biological control: New challenges in an old relationship. *Neotropical Entomology*, page 22, 2023. doi: 10.1007/s13744-023-01025-5. 24

Ali Shojaie and Emily B Fox. Granger causality: A review and recent advances. *Annual Review of Statistics and Its Application*, 9:289 – 319, 2022. 109

Fábio Amaral Godoy da Silveira, Everton Castelão Tetila, Gilberto Astolfi, Anderson Bessa da Costa, and Willian Paraguassu Amorim. Performance analysis of YOLOv3 for real-time detection of pests in soybeans. In *Brazilian Conference on Intelligent Systems*, pages 265 – 279. Springer, 2021. 26

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv 1409.1556*, 09 2014. 18, 69

Amit Singh, Rakesh Kumar Dwivedi, and Rajul Rastogi. Machine learning based framework for lung cancer detection and image feature extraction using vgg16 with pca on ct-scan images. *SN Computer Science*, 5, 2024a. 69

Ankit Kumar Singh, Md Yeasin, Ranjit Kumar Paul, AK Paul, and Anita Sarkar. Dynamic ensemble-based machine learning models for predicting pest populations. *Frontiers in Applied Mathematics and Statistics*, 10:1435517, 2024b. 8

SO Slim, IA Abdelnaby, MS Moustafa, MB Zahran, HF Dahi, and MS Yones. Smart insect monitoring based on YOLOv5 case study: Mediterranean fruit fly ceratitis capitata and peach fruit fly bactrocera zonata. *The Egyptian Journal of Remote Sensing and Space Sciences*, 26:881 – 891, 2023. 67

IN Smyrnioudis, R Harrington, SJ Clark, and N Katis. The effect of natural enemies on the spread of barley yellow dwarf virus (bydv) by rhopalosiphum padi (hemiptera: Aphididae). *Bulletin of Entomological Research*, 91:301 – 306, 2001. 106

LO Solis-Sánchez, JJ García-Escalante, R Castañeda-Miranda, I Torres-Pacheco, and R Guevara-González. Machine vision algorithm for whiteflies (bemisia tabaci genn.) scouting under greenhouse environment. *Journal of applied entomology*, 133:546 – 552, 2009. 25

Witenberg SR Souza, Adão Nunes Alves, and Dıbio Leandro Borges. A deep learning model for recognition of pest insects in maize plantations. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pages 2285 – 2290. IEEE, 2019. 25

Evangelos Spiliotis, Spyros Makridakis, Artemios-Anargyros Semenoglou, and Vassilios Assimakopoulos. Comparison of statistical and machine learning methods for daily sku demand forecasting. *Operational Research*, pages 1 – 25, 2020. 107

Elias Sprengel, Martin Jaggi, Yannic Kilcher, and Thomas Hofmann. Audio based bird species identification using deep learning techniques. *LifeCLEF 2016*, pages 547 – 559, 2016. 16

Amrita Srivathsan, Vivian Feng, Daniel Suárez, Brent Emerson, and Rudolf Meier. Ontbarcoder 2.0: rapid species discovery and identification with real-time barcoding facilitated by oxford nanopore r10. 4. *Cladistics*, 40:192 – 203, 2024. 66

C Srujana, B Sriya, S Divya, Subhani Shaik, and V Kakulapati. Species identification of birds via acoustic processing signals using recurrent network analysis (rnn). In *International Conference on Soft Computing and Signal Processing*, pages 27 – 38. Springer, 2023. 7

Dirk Steinke, Sujeevan Ratnasingham, Jireh Agda, Hamzah Ait Boutou, Isaiah CH Box, Mary Boyle, Dean Chan, Corey Feng, Scott C Lowe, Jaclyn TA McKeown, et al. Towards a taxonomy machine: A training set of 5.6 million arthropod images. *Data*, 9, 2024. 5, 6

Johan A Stenberg, Ingvar Sundh, Paul G Becher, Christer Björkman, Mukesh Dubey, Paul A Egan, Hanna Friberg, José F Gil, Dan F Jensen, Mattias Jonsson, et al. When is it biological control? a framework of definitions, mechanisms, and classifications. *Journal of Pest Science*, 94:665 – 676, 2021. 24

VMRF Stern, R Smith, Robert van den Bosch, Kenneth Hagen, et al. The integration of chemical and biological control of the spotted alfalfa aphid: the integrated control concept. *Hilgardia*, 29:81 – 101, 1959. 87

Rainer Storn and Kenneth Price. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization*, 11:341 – 359, 1997. 138, 139

Thomas Stützle, Marco Dorigo, et al. Aco algorithms for the traveling salesman problem. *Evolutionary algorithms in engineering and computer science*, 4:163 – 183, 1999. 71

Mukilan Deivarajan Suresh, Tong Xin, Samantha M Cook, and Darren M Evans. Bugs and bytes: Entomological biomonitoring through the integration of deep learning and molecular analysis for merged community and network analysis. *Agricultural and Forest Entomology*, 27:35 – 49, 2025. 5, 6

Ankita Suryavanshi, Vinay Kukreja, and Rajat Saini. Feathered insights: Advanced cnn-based classification of sparrow species. In *2024 5th IEEE Global Conference for Advancement in Technology (GCAT)*, pages 1 – 5. IEEE, 2024. 7

Richard Szeliski. *Computer vision: algorithms and applications*. Springer Nature, 2022. 26, 29

Patrick L Taggart, Brian Cooke, David E Peacock, Peter West, Emma Sawyers, and Kandarp K Patel. Do land managers apply best-practice integrated pest management: a case study of the European rabbit. *Journal of Pest Science*, pages 1 – 16, 2024. 107

Alia Tajdar, Chuan Cao, Khalid Abbas, Muhammad Shah Zaib, Hafiz Muhammad Safeer, Syed Muhammad Zaka, Wangpeng Shi, and Waqar Jaleel. Monitoring activity of spodoptera frugiperda (smith) in different areas of maize crops and its pesticide susceptibility testing under controlled conditions. *Journal of Toxicology*, 2025:6651151, 2025. 8

Pejman Tajmiri, Seyed Ali Asghar Fathi, Ali Golizadeh, and Gadir Nouri-Ganbalani. Effect of strip-intercropping potato and annual alfalfa on populations of leptinotarsa decemlineata say and its predators. *International Journal of Pest Management*, 63:273 – 279, 2017. 7

Jun Tang, Gang Liu, and Qingtao Pan. A review on representative swarm intelligence algorithms for solving optimization problems: Applications and trends. *IEEE/CAA Journal of Automatica Sinica*, 8:1627 – 1643, 2021. 71

Jing Tao, Min Chen, Shixiang Zong, and You-Qing Luo. Genetic structure in the seabuckthorn carpenter moth (holcocerus hippophaecolus) in China: The role of outbreak events, geographical and host factors. *PloS one*, 7, 01 2012. doi: 10.1371/journal.pone.0030544. 86

Haruki Tatsuta, Kazuo H Takahashi, and Yositaka Sakamaki. Geometric morphometrics in entomology: Basics and applications. *Entomological Science*, 21:164 – 184, 2018. 26

Ana Cláudia Teixeira, José Ribeiro, Raul Morais, Joaquim J. Sousa, and António Cunha. A systematic review on automatic insect detection using deep learning. *Agriculture*, 13, 2023. ISSN 2077-0472. doi: 10.3390/agriculture13030713. URL https://www.mdpi.com/2077-0472/13/3/713. 4, 5

Everton Castelão Tetila, Bruno Brandoli Machado, Geazy Vilharva Menezes, Nicolas Alessandro de Souza Belete, Gilberto Astolfi, and Hemerson Pistori. A deep-learning approach for automatic counting of soybean insect pests. *IEEE Geoscience and Remote Sensing Letters*, 17:1837 – 1841, 2019. 25, 26

Sarjak Thakkar, Changxing Cao, Lifan Wang, Tae Jong Choi, and Julian Togelius. Autoencoder and evolutionary algorithm for level generation in lode runner. In *2019 IEEE Conference on Games (CoG)*, pages 1 – 4. IEEE, 2019. 71

K Thenmozhi and U Srinivasulu Reddy. Crop pest classification based on deep convolutional neural network and transfer learning. *Computers and Electronics in Agriculture*, 164:104906, 2019. 25, 26

Lori Ann Thrupp. The importance of biodiversity in agroecosystems. *Journal of Crop Improvement*, 12:315 – 337, 2004. 65

N. Tinsley, Ronald Estes, and M. Gray. Validation of a nested error component model to estimate damage caused by corn rootworm larvae. *Journal of Applied Entomology*, 137, 04 2013. doi: 10.1111/j.1439-0418.2012.01736.x. 87

Junlong Tong, Liping Xie, Wankou Yang, Kanjian Zhang, and Junsheng Zhao. Enhancing time series forecasting: a hierarchical transformer with probabilistic decomposition representation. *Information Sciences*, 647, 2023. 129

R Toscano-Miranda, M Toro, J Aguilar, M Caro, A Marulanda, and A Trebilcok. Artificial-intelligence and sensing techniques for the management of insect pests and diseases in cotton: a systematic literature review. *The Journal of Agricultural Science*, 160:16 – 31, 2022. 26

Constantino Tsallis. Possible generalization of boltzmann-gibbs statistics. *Journal of statistical physics*, 52:479 – 487, 1988. 94

Constantino Tsallis and Daniel A Stariolo. Generalized simulated annealing. *Physica A: Statistical Mechanics and its Applications*, 233:395 – 406, 1996. 75, 94, 138, 139

Ilias Tsoumas, Vasileios Sitokonstantinou, Georgios Giannarakis, Evagelia Lampiri, Christos Athanassiou, Gustau Camps-Valls, Charalampos Kontoes, and Ioannis N Athanasiadis. Causality and explainability for trustworthy integrated pest management. In *NeurIPS 2023 Workshop on Tackling Climate Change with Machine Learning*, 2023. URL https://www.climatechange.ai/papers/neurips2023/119. 8

Ilias Tsoumas, Vasileios Sitokonstantinou, Georgios Giannarakis, Evagelia Lampiri, Christos Athanassiou, Gustau Camps-Valls, Charalampos Kontoes, and Ioannis N. Athanasiadis. Leveraging causality and explainability in digital agriculture. *Environmental Data Science*, 4:e23, 2025. doi: 10.1017/eds.2025.14. 8

Kenta Uchida, Rachel V Blakey, Joseph R Burger, Daniel S Cooper, Chase A Niesner, and Daniel T Blumstein. Urban biodiversity and the importance of scale. *Trends in Ecology & Evolution*, 36:123 – 131, 2021. 65

Muhib Ullah, Muhammad Shabbir Hasan, Abdul Bais, Tyler Wist, and Shaun Sharpe. A novel computer vision system for efficient flea beetle monitoring in canola crop. *IEEE Transactions on AgriFood Electronics*, 2024. 5

Abhishek Upadhyay, Narendra Singh Chandel, Krishna Pratap Singh, Subir Kumar Chakraborty, Balaji M Nandede, Mohit Kumar, A Subeesh, Konga Upendar, Ali Salem, and Ahmed Elbeltagi. Deep learning and computer vision in plant disease detection: a comprehensive review of techniques, models, and trends in precision agriculture. *Artificial Intelligence Review*, 58:1 – 64, 2025. 3

Gopi Upreti. Importance of biodiversity, ecosystems, and ecosystem services. In *Ecosociocentrism: The Earth First Paradigm for Sustainable Living*, pages 15 – 30. Springer, 2023. 65

Roozbeh Valavi, Jane Elith, José J Lahoz-Monfort, and Gurutzeta Guillera-Arroita. Modelling species presence-only data with random forests. *Ecography*, 44:1731 – 1742, 2021. 115

Roel Van Klink, Tom August, Yves Bas, Paul Bodesheim, Aletta Bonn, Frode Fossøy, Toke T Høye, Eelke Jongejans, Myles HM Menz, Andreia Miraldo, et al. Emerging technologies revolutionise insect ecology and monitoring. *Trends in ecology & evolution*, 37:872 – 885, 2022. 3, 4, 66

Roel Van Klink, Julie Koch Sheard, Toke T Høye, Tomas Roslin, Leandro A Do Nascimento, and Silke Bauer. Towards a toolkit for global insect biodiversity monitoring, 2024. 1, 2

Simon van Noort. The role of taxonomy and museums in insect conservation. In *Routledge Handbook of Insect Conservation*, pages 450 – 460. Taylor & Francis Group, 2024. 65

Guido Van Rossum and Fred L Drake Jr. *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995. 98

Derek van Tilborg, Helena Brinkmann, Emanuele Criscuolo, Luke Rossen, Rıza Özçelik, and Francesca Grisoni. Deep learning for low-data drug discovery: hurdles and opportunities. *Current Opinion in Structural Biology*, 86, 2024. 66

Brian W van Wilgen, S Raghu, Andy W Sheppard, and Urs Schaffner. Quantifying the social and economic benefits of the biological control of invasive alien plants in natural ecosystems. *Current opinion in insect science*, 38:1 – 5, 2020. 24

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 74, 129

Stephanie Vaz, Stella Manes, Gabriel Khattar, Mariana Mendes, Luiz Silveira, Eduardo Mendes, Erimagna de Morais Rodrigues, Danielle Gama-Maia, Maria Lucia Lorini, Margarete Macedo, et al. Global meta-analysis of urbanization stressors on insect abundance, richness, and traits. *Science of the Total Environment*, 2023. 65

Titus Venverloo and Fábio Duarte. Towards real-time monitoring of insect species populations. *Scientific Reports*, 14, 2024. 56

Luciano M Verdade, Maria Carolina Lyra-Jorge, Carlos I Piña, et al. *Applied ecology and human dimensions in biological conservation*. Springer, 2014. 1, 2

W.Jan Volney and Richard Fleming. Climate change and impacts of boreal forest insects. *Agriculture, Ecosystems and Environment*, 82:283 – 294, 12 2000. doi: 10.1016/S0167-8809(00)00232-2. 86

J Wolfgang Wägele, Paul Bodesheim, Sarah J Bourlat, Joachim Denzler, Michael Diepenbroek, Vera Fonseca, Karl-Heinz Frommolt, Matthias F Geiger, Birgit Gemeinholzer, Frank Oliver Glöckner, et al. Towards a multisensor station for automated biodiversity monitoring. *Basic and Applied Ecology*, 59:105 – 138, 2022. 1

David L Wagner, Eliza M Grames, Matthew L Forister, May R Berenbaum, and David Stopak. Insect decline in the anthropocene: Death by a thousand cuts. *Proceedings of the National Academy of Sciences*, 118, 2021. 65

W Wallner. Factors affecting insect population dynamics: Differences between outbreak and non-outbreak species. *Annual Review of Entomology*, 32:317 – 340, 11 1987. doi: 10.1146/annurev.en.32.010187.001533. 85, 86

Bo Wang. Identification of crop diseases and insect pests based on deep learning. *Scientific Programming*, 2022:1 – 10, 2022. 26

Jiale Wang, Yan Chen, Jianxiang Huang, Xunyuan Jiang, and Kai Wan. Leveraging machine learning for advancing insect pest control: A bibliometric analysis. *Journal of Applied Entomology*, 2024a. 4, 5

Jinfeng Wang, Yunqiang Chen, Kaihong Zheng, Zhipeng Cheng, Renyou Yang, and Qiong Huang. Fcls: A lightweight model based on sound feature fusion for bird classification. In *2024 International Conference on Virtual Reality and Visualization (ICVRV)*, pages 30–35, 2024b. doi: 10.1109/ICVRV62410.2024.00015. 7

Rick Wang, Amir-Hossein Karimi, and Ali Ghodsi. Distance correlation autoencoder. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1 – 8. IEEE, 2018. 74

Wenlin Wang, Zhe Gan, Hongteng Xu, Ruiyi Zhang, Guoyin Wang, Dinghan Shen, Changyou Chen, and Lawrence Carin. Topic-guided variational autoencoders for text generation. *arXiv preprint arXiv:1903.07137*, 2019. 40, 44

Xuquan Wang, Zhiyuan Ma, Yujie Xing, Tianfan Peng, Xiong Dun, Zhuqing He, Jian Zhang, and Xinbin Cheng. Rapid species discrimination of similar insects using hyperspectral imaging and lightweight edge artificial intelligence. *Royal Society Open Science*, 11:240485, 2024c. 5, 6

Yingfan Wang, Haiyang Huang, Cynthia Rudin, and Yaron Shaposhnik. Understanding how dimension reduction tools work: An empirical approach to deciphering t-sne, umap, trimap, and pacmap for data visualization. *J. Mach. Learn. Res.*, 22:1 – 73, 2021. 26, 47, 48

M Arif Wani, Farooq Ahmad Bhat, Saduf Afzal, and Asif Iqbal Khan. *Advances in deep learning*. Springer, 2020. 18, 25, 26, 44

Yu-Cheng Wei, Wei-Lun Chen, Mao-Ning Tuanmu, Sheng-Shan Lu, and Ming-Tang Shiao. Advanced montane bird monitoring using self-supervised learning and transformer on passive acoustic data. *Ecological Informatics*, 84:102927, 2024. 7

Dominika Winiarska, Paweł Szymański, and Tomasz S Osiejuk. Detection ranges of forest bird vocalisations: guidelines for passive acoustic monitoring. *Scientific Reports*, 14:894, 2024. 6

Tarid Wongvorachan, Surina He, and Okan Bulut. A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining. *Information*, 14, 2023. 67

Lichao Wu and Stjepan Picek. Remove some noise: On pre-processing of side-channel measurements with autoencoders. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, pages 389 – 415, 2020. 74

Y Xiang and XG Gong. Efficiency of generalized simulated annealing. *Physical Review E*, 62, 2000. 94

Y Xiang, DY Sun, W Fan, and XG Gong. Generalized simulated annealing algorithm and its application to the thomson model. *Physics Letters A*, 233:216 – 220, 1997. 94

Yang Xiang, Sylvain Gubian, Brian Suomela, and Julia Hoeng. Generalized simulated annealing for global optimization: the gensa package. *R J.*, 5, 2013. 94

Zeyu Xu, Tiejun Wang, Andrew K Skidmore, and Richard Lamprey. A review of deep learning techniques for detecting animals in aerial and satellite images. *International Journal of Applied Earth Observation and Geoinformation*, 128: 103732, 2024. 3

Huiyong Yang, Wei Liu, Kun Xing, Jian Qiao, Xin Wang, Lingwang Gao, and Zuorui Shen. Research on insect identification based on pattern recognition technology. In *2010 Sixth International Conference on Natural Computation*, volume 2, pages 545 – 548. IEEE, 2010. 25

Han Gyu Yoon, Chanki Lee, Doo Bong Lee, Seung Min Park, Jun Woo Choi, Hee Young Kwon, and Changyeon Won. Interpolation and extrapolation between the magnetic chiral states using autoencoder. *Computer Physics Communications*, 272, 2022. 73

Alex Eric Yuan and Wenying Shou. Data-driven causal analysis of observational biological time series. *Elife*, 11, 2022. 129

Yage Yuan, Jianan Wei, Haisong Huang, Weidong Jiao, Jiaxin Wang, and Hualin Chen. Review of resampling techniques for the treatment of imbalanced industrial data classification in equipment condition monitoring. *Engineering Applications of Artificial Intelligence*, 126, 2023. 67

Mohd Yusri Zainudin, Saiful Zaimi Jamil, Mohd Syauqi Nazmi, and Mohd Fuad Mohd Nor. Effects of integrated pest management (ipm) practice on in-

sects population and yield of cabbage in cameron highlands. *AgroTech-Food Science, Technology and Environment*, 2:7 – 13, 2023. 7

Teresinha Zanuncio, José Zanuncio, Fernando Freitas, Dirceu Pratissoli, Camilla Sediyama, and Vanessa Maffia. Main lepidopteran pest species from an eucalyptus plantation in minas gerais, Brazil. *Revista de biología tropical*, 54:553 – 60, 07 2006. doi: 10.15517/rbt.v54i2.13922. 86

IY Zayas and Paul W Flinn. Detection of insects in bulkwheat samples with machine vision. *Transactions of the ASAE*, 41:883, 1998. 25

Mingfeng Zha, Wenbin Qian, Wenlong Yi, and Jing Hua. A lightweight YOLOv4-based forestry pest detection method using coordinate attention and feature fusion. *Entropy*, 23, 2021. 26

Aston Zhang, Zachary C. Lipton, Mu Li, and Alexander J. Smola. Dive into deep learning. *arXiv preprint arXiv:2106.11342*, 2021. 36, 37, 40

Aston Zhang, Zachary C. Lipton, Mu Li, and Alexander J. Smola. *Dive into Deep Learning*. Cambridge University Press, 2023a. https://D2L.ai. 74

Haowen Zhang, Shengyuan Zhao, Yifei Song, Shishuai Ge, Dazhong Liu, Xianming Yang, and Kongming Wu. A deep learning and grad-cam-based approach for accurate identification of the fall armyworm (spodoptera frugiperda) in maize fields. *Computers and Electronics in Agriculture*, 202, 2022. 26, 44

Jinglan Zhang, Kai Huang, Mark Cottman-Fields, Anthony Truskinger, Paul Roe, Shufei Duan, Xueyan Dong, Michael Towsey, and Jason Wimmer. Managing and analysing big audio data for environmental monitoring. In *2013 IEEE 16th International Conference on Computational Science and Engineering*, pages 997 – 1004. IEEE, 2013. 6

Lexin Zhang, Kuiheng Chen, Liping Zheng, Xuwei Liao, Feiyu Lu, Yilun Li, Yuzhuo Cui, Yaze Wu, Yihong Song, and Shuo Yan. Enhancing fruit fly detection in complex backgrounds using transformer architecture with step attention mechanism. *Agriculture*, 14, 2024. 67

Pengyuan Zhang, Hangting Chen, Haichuan Bai, and Qingsheng Yuan. Deep scattering spectra with deep neural networks for acoustic scene classification tasks. *Chinese Journal of Electronics*, 28:1177 – 1183, 2019. 16

Xiaolei Zhang, Junyi Bu, Xixiang Zhou, and Xiaochan Wang. Automatic pest identification system in the greenhouse based on deep learning and machine vision. *Frontiers in Plant Science*, 14, 2023b. 55

Naizhuo Zhao, Katia Charland, Mabel Carabali, Elaine O Nsoesie, Mathieu Maheu-Giroux, Erin Rees, Mengru Yuan, Cesar Garcia Balaguera, Gloria Jaramillo Ramirez, and Kate Zinszer. Machine learning and dengue forecasting: Comparing random forests and artificial neural networks for predicting dengue burden at national and sub-national scales in colombia. *PLOS Neglected Tropical Diseases*, 14, 2020. 107

Nan Zhao, Lei Zhou, Ting Huang, Mohamed Farag Taha, Yong He, and Zhengjun Qiu. Development of an automatic pest monitoring system using a deep learning model of dpenet. *Measurement*, 203:111970, 2022. 2

Shulin Zhao, Xiaoting Sun, and Lingyun Gai. Data enhancement and multi-feature learning model for pest classification. *Journal of Intelligent & Fuzzy Systems*, 45:5409 – 5421, 2023. 6

Li Zhigang, Fu Zetian, Shi Yan, and Xia Tiehua. Prototype system of automatic identification cotton insect pests and intelligent decision based on machine vision. In *2003 ASAE Annual Meeting*, page 1. American Society of Agricultural and Biological Engineers, 2003. 25

Ningxing Zhou, Tyler Wist, and Sean M. Prager. Economic thresholds and economic injury level for pea aphid in tannin and low tannin faba bean. *Crop Protection*, 186:106919, 2024. ISSN 0261-2194. doi: https://doi.org/10.1016/j.cropro.2024.106919. URL https://www.sciencedirect.com/science/article/pii/S0261219424003478. 8

Boaz Zion. The use of computer vision technologies in aquaculture–a review. *Computers and electronics in agriculture*, 88:125 – 132, 2012. 3

Elise F. Zipkin and Jeffrey W. Doser. Context matters in ecological forecasting: Lessons in predicting species distributions. *Global Change Biology*, 30:e17123, January 2024. ISSN 1354-1013, 1365-2486. doi: 10.1111/gcb.17123. URL https://onlinelibrary.wiley.com/doi/10.1111/gcb.17123. 8

Valentin Ștefan, Thomas Stark, Michael Wurm, Hannes Taubenböck, and Tiffany M Knight. Successes and limitations of pretrained YOLO detectors applied to unseen time-lapse images for automated pollinator monitoring. *Scientific Reports*, 15, 2025. 55