EXPLAINABLE MACHINE LEARNING TECHNIQUES FOR APPLICATIONS IN LIFE SCIENCES

Martin Marzidovšek

Doctoral Dissertation
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

Supervisor: Dr. Vid Podpečan, Jožef Stefan Institute, Ljubljana, Slovenia
Co-Supervisor: Prof. Dr. Patricija Mozetič, National Institute of Biology, Ljubljana, Slovenia

Evaluation Board:
Prof. Dr. Nada Lavrač, Chair, Jožef Stefan Institute, Ljubljana, Slovenia
Prof. Dr. Ljupčo Todorovski, Member, Faculty of Mathematics and Physics, University of Ljubljana, Ljubljana, Slovenia
Prof. Dr. Christian Mulder, Member, University of Catania, Catania, Italy

MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA
JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL

Martin Marzidovšek

EXPLAINABLE MACHINE LEARNING TECHNIQUES FOR APPLICATIONS IN LIFE SCIENCES

Doctoral Dissertation

RAZLOŽLJIVE TEHNIKE STROJNEGA UČENJA V ZNANOSTIH O ŽIVLJENJU

Doktorska disertacija

Supervisor: Dr. Vid Podpečan

Co-Supervisor: Prof. Dr. Patricija Mozetič

Ljubljana, Slovenia, September 2024

To my parents and Mina
Mojim staršem in Mini

Acknowledgments

I would especially like to thank my supervisor Vid Podpečan for his support and guidance in coding, general directions for research and writing papers, as well as for his trust and participation in all our research projects.

I thank my co-supervisor Patricija Mozetič for her help in writing the papers and understanding topics in marine biology.

I thank Nada Lavrač and Sašo Džeroski for giving me the opportunity to work at the Jožef Stefan Institute.

I thank Mili Bauer for her help throughout my PhD studies and beyond.

I thank Matej Petkovič for his technical advice during the writing of the dissertation.

Finally, I thank the evaluation committee – Nada Lavrač, Christian Mulder and Ljupčo Todorovski – for taking the time and for corrections that improved the dissertation.

Many thanks to all of you.

I acknowledge the following funding grants: Horizon 2020 project EMBEDDIA (ID: 825153) and ARIS (J4-2544, J3-3070, J7-4636).

Abstract

As ecological, agricultural, and biological disciplines face mounting challenges like biodiversity loss, food chain disruption, and climate change, leveraging machine learning (ML) to process complex and heterogeneous data becomes increasingly vital. This dissertation explores the potential of ML in combination with explainability approaches for enhancing research in life sciences, specifically focusing on ecology. The research aims to advance ML applications in ecosystem monitoring, ecological modelling, and predictive analytics, emphasising the necessity for explainable artificial intelligence (xAI) to balance predictive performance with model transparency.

The dissertation underscores the potential for xAI and interpretable ML models. While advanced ML models excel in handling non-linear relationships and complex systems, their 'black box' nature often limits their scientific utility. This work advocates for the integration of xAI methods to elucidate the inner workings of these models, thereby enhancing their applicability and acceptance in scientific research.

The contributions of the dissertation include:

- Modelling diarrhetic shellfish poisoning events: An xAI approach predicting the toxicity of mussels due to harmful algal blooms in the Adriatic Sea. Utilising a 28-year dataset, the study highlights the importance of data pre-processing and demonstrated that Random Forest models, coupled with explainability methods, provide critical insights into marine ecosystem dynamics and can serve as cost-effective early warning systems.

- Unsupervised ML in Agriculture: Cluster analysis was applied to identify barriers and incentives for the use of decision support systems (DSS) in integrated pest management. The study revealed distinct groups among farmers and advisors, identifying common barriers and incentives, which can inform future research to enhance DSS adoption.

- Phytoplankton Identification and Quantification: A computer vision system that automates the identification, size estimation, and biovolume calculation of phytoplankton species. Using transfer learning, the system processes samples more efficiently than manual methods, providing a more accurate assessment of marine ecosystems. Visual explanations are applied for greater thrustworthiness and confidence of the ML-based solution.

- BEFANA Software: An open-source software tool that facilitates network analysis and ML applications in ecological networks. It enables ecologists interactive visualisation and to quantify network topologies, test hypotheses, and embed experimental data, thereby enriching the analysis and modelling of ecosystem dynamics.

The dissertation contributes to improving ML methodologies in life sciences research. The principles of open, reproducible science are upheld through open access to software code, data, and publications associated with this research.

Povzetek

Ekološke, kmetijske in biološke discipline se soočajo z vse večjimi izzivi, kot so izguba biotske raznovrstnosti, motnje v prehranjevalnih verigah in podnebne spremembe, zato postaja uporaba strojnega učenja za obdelavo kompleksnih in heterogenih podatkov vse bolj pomembna. Disertacija raziskuje potencial strojnega učenja v kombinaciji z razlagalnimi pristopi za izboljšanje raziskav v znanostih o življenju, s posebnim poudarkom na ekologiji. Raziskava si prizadeva doprinesti k uporabi strojnega učenja na področju spremljanja ekosistemov, ekološkega modeliranja in napovednih analiz, pri čemer poudarja potrebo po razložljivi umetni inteligenci, da se bolje uravnoteži napovedno zmogljivost z razložljivostjo modela.

Disertacija poudarja potencial razložljivih modelov strojnega učenja. Čeprav napredni modeli strojnega učenja lahko zajamejo nelinearne odnose in kompleksne sisteme, je njihova uporabnost v znanosti omejena zaradi njihove slabše razložljivosti. Delo zagovarja uporabo metod razložljive umetne inteligence za pojasnitev notranjega delovanja teh modelov in s tem povečuje njihovo uporabnost in sprejemljivost v znanstvenih raziskavah.

Prispevki disertacije vključujejo:

- Modeliranje pojavnosti toksičnosti školjk: Pristop razložljive umetne inteligence za napovedovanje toksičnosti školjk zaradi škodljivega cvetenja alg v Jadranskem morju. Z uporabo 28-letnega nabora podatkov študija poudarja pomembnost predobdelave podatkov in dokazuje, da modeli naključnega gozda v kombinaciji z metodami razložljivosti zagotavljajo ključne vpoglede v dinamiko morskih ekosistemov in lahko služijo kot stroškovno učinkovit opozorilni sistem.

- Nenadzorovano strojno učenje v kmetijstvu: Uporaba gručenja za identifikacijo ovir in spodbud za uporabo sistemov za podporo odločanja pri integriranem varstvu rastlin. Študija je odkrila različne skupine med kmeti in svetovalci, pri čemer so bile identificirane skupne ovire in spodbude, ki lahko usmerjajo prihodnje raziskave za izboljšanje uporabe sistemov za podporo odločanja.

- Identifikacija in kvantifikacija fitoplanktona: Sistem računalniškega vida za avtomatizirano identifikacijo, oceno velikosti in izračun biovolumna vrst fitoplanktona. Z uporabo učenja s prenosom znanja sistem obdeluje vzorce učinkoviteje kot ročne metode, kar omogoča natančnejšo analizo morskih ekosistemov. Vizualne razlage so uporabljene za izboljšanje zaupanja in zanesljivosti rešitev strojnega učenja.

- Programska oprema BEFANA: Odprtokodno orodje za analizo omrežij in uporabe strojnega učenja v ekoloških omrežjih. Ekologom omogoča interaktivno vizualizacijo, kvantifikacijo topologij omrežij, preizkušanje hipotez in vključevanje eksperimentalnih podatkov, kar obogati analizo in modeliranje dinamike ekosistemov.

Disertacija prispeva k izboljšanju uporabe metod strojnega učenja v raziskavah v znanostih o življenju. Načela odprte in ponovljive znanosti so upoštevana z odprtim dostopom do programske kode, podatkov in publikacij, povezanih s to raziskavo.

Contents

List of Figures

Abbreviations

AI	artificial intelligence
ANN	artificial neural network
BRT	boosted regression trees
CNN	convolutional neural network
COCO	Common objects in context
DBCV	density-based clustering validation
DL	deep learning
DSS	decision support system
DT	decision tree
E	experience
EWS	early warning systems
GNN	graph neural network
HAB	harmful algal bloom
IPM	integrated pest management
IoU	intersection over union
ML	machine learning
MLP	multilayer perceptron
NN	neural network
P	performance
RF	random forest
SVM	support vector machines
T	task
UMAP	Uniform manifold approximation and projection
YOLO	You only look once
xAI	explainable artificial intelligence
mAP	mean average precision

Chapter 1

Introduction

This chapter presents the background of the relevant topics, the motivation and goals of the dissertation. It presents the goals and research hypotheses, and provides a brief overview of the scientific contributions, which are presented in more detail in the subsequent chapters. Finally, the structure of the dissertation is outlined.

1.1 Background

Machine learning (ML), a branch of artificial intelligence (AI), offers versatile and powerful tools for analysing data and predictive modelling. Mitchel (1997) provides a concise definition of the discipline: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E". ML aims to automatically extract information (knowledge) from data (E) and discover regularities that enable the construction of a general and accurate model that can make predictions for unseen data (Cios et al., 2007). The main use case of ML is dealing with prediction problems, the mapping of inputs to outputs. This involves discerning patterns in the data in order to model complex predictive relationships based on statistical inference from the data (Hastie et al., 2009).

ML follows different learning paradigms such as supervised learning, unsupervised learning and reinforcement learning. In the scientific field, ML is used for prediction problems, understanding the mechanisms underlying these predictions and handling complex transformations of input data (Raghu & Schmidt, 2020). The surge of high-dimensional data in science often requires dimensionality reduction and pattern recognition. Unsupervised learning algorithms handle this effectively before the data is used in downstream supervised learning tasks (Pichler & Hartig, 2023).

Over the last decade, the development of deep learning (DL), a subset of ML models based on artificial neural networks (ANN), has enabled significant progress in numerous ML tasks (Lecun et al., 2015). The emergence of DL is attributed to the co-evolution of computational resources, ML algorithms and data availability. DL has revolutionised areas from speech recognition to image classification and natural language processing (Raghu & Schmidt, 2020) and many of these fundamental tasks have the potential for investigating scientific questions.

1.2 Machine Learning in Life Sciences

In recent years, ML, AI and DL have revolutionised almost all areas of science (Jordan & Mitchell, 2015). This work takes a closer look at life sciences, a branch of natural sciences that deals with the study of biological systems – from microorganisms, plants, animals to humans – and in which ML methods are increasingly used. The main reasons for this are the exponential increase in the amount of data thanks to advanced sensors and the greater capacity of ML algorithms to cope with the volume and complexity of biological data (Heil et al., 2021).

The life sciences are characterised by an era of vast amounts of data of varied modalities, including numerical, textual, image, sound, video and graph structured data. An important category of data generated in life science research is molecular data, which is expressed in the form of nucleotides, amino acids, DNA, RNA, peptides and proteins and requires efficient processing (Cios et al., 2007). ML methods have the unique ability to manage large amounts of heterogeneous data, even in unstructured form. They also excel in examining biological systems without imposing unrealistic assumptions. They are able to infer missing data and can reduce the long-term effort for annotation by experts (Thessen, 2016).

ML has found numerous applications in the sub-disciplines of the life sciences and promoted important scientific discoveries. ML and AI offer novel solutions with the potential to impact health, agriculture, pharmaceutics and more, contributing to many of the Sustainable Development Goals (United Nations, 2015). Some of the most impactful examples include food security, clean water, climate action, sustainable use of terrestrial ecosystems, combating desertification and reversing land degradation and biodiversity loss (Holzinger et al., 2023).

This work focuses largely on specific ML applications in ecology – an important subdiscipline of life sciences that studies organisms in their environmental context and is essential for addressing the above challenges. In ecological research, recent advances in observational technology have facilitated the collection of large-scale, multi-source data (Guo et al., 2020). However, rapid and large-scale interpretation of such data for ecological insights remains a challenge (Tuia et al., 2022). Therefore, the application of ML methods is crucial for the advancement of ecological research.

By bridging the gap between the vast amounts of available data and the need for extracting actionable insights, ML has proven instrumental in ecosystem monitoring and ecological modelling (Christin et al., 2019; Kwok, 2019). Comprehensive research, understanding of ecological resources and predicting ecosystem evolution facilitated by ML are essential for the sustainable development of human society (Guo et al., 2020).

Advances in ML methodology are accelerating different phases of traditional ecological research, from targeted data collection to semi-automated population surveys. These developments enable more comprehensive studies of animal abundances, diversity and behaviour (Tuia et al., 2022). In some cases, ML can significantly accelerate ecological research and also reduce analysis costs with reduction factors estimated at 2 to 10 (Eikelboom et al., 2019).

The benefits of big data and ML are increasingly recognised in ecological research and awareness of ML methods among domain experts is growing (Christin et al., 2019; Farley et al., 2018; Peters et al., 2014; Tuia et al., 2022; Weinstein, 2018). ML methods are frequently used in ecology for the identification of species from images, videos and sounds (Mac Aodha et al., 2018; Tabak et al., 2020; Wäldchen & Mäder, 2018) and the modelling of species distribution (Debeljak & Džeroski, 2011; Dzeroski & Todorovski, 2007; Elith et al., 2006; Gobeyn et al., 2019).

Graph theory is also a useful approach for answering ecological questions related to the structure of an ecological network (Dale, 2017). ML on graphs has also been applied in ecology for important problems such as filling missing links in ecological networks (Desjardins-Proulx et al., 2017). Techniques such as graph neural networks (GNN), which work directly on the nodes and their edges, are used for tasks such as classifying nodes or edges and link prediction.

While ML's ability to predict patterns is obvious, it holds even greater potential for ecological inference by promoting understanding of the processes underlying these patterns (Lucas, 2020). However, its use for causal inference is still controversial, as the focus of ML is on prediction rather than model interpretability (as will be shown later). It is expected that these methods will become an indispensable tool in ecology and evolution, comparable to other traditional statistical tools (Pichler & Hartig, 2023).

1.2.1 Challenges Relating to Data

In contrast to some other subfields of life sciences where data is abundant, one of the biggest challenges in applying modern ML algorithms in ecology is the scarcity of large datasets needed for model training. In ecology, small datasets are not uncommon as observations can be difficult or costly to obtain. Newer ML techniques such as few-shot learning (Wang et al., 2020) and transfer learning can significantly reduce the amount of data required, making them particularly advantageous for ecologists (Pichler & Hartig, 2023). However, these techniques are primarily suitable for image processing-based tasks and deep learning, not for classic structured data (Pichler & Hartig, 2023).

The nature of climate-relevant data presents unique challenges and opportunities. The data may be proprietary, contain sensitive personal information or come from heterogeneous sources, which requires integration using domain knowledge (Rolnick et al., 2022). In addition, combining these datasets can be complicated due to variations in ecological patterns across scales (Maglianesi et al., 2014; Poisot et al., 2015; Strydom et al., 2021). Existing datasets may not be organised with a specific task in mind, unlike typical ML benchmarks that have a clear objective (Rolnick et al., 2022).

Ecological data can also exhibit low signal-to-noise ratio, missing values and imbalanced classes (for classification tasks), which can make applications of ML more difficult, as evidenced by the published research presented below.

1.2.2 Prediction Performance vs. Understanding

DL algorithms have indeed demonstrated exceptional capabilities in dealing with image-based tasks and complex, unstructured data. Nevertheless, classical ML algorithms such as Random Forest (RF) and Boosted Regression Trees (BRT) still outperform them in the field of structured data (Pichler & Hartig, 2023). It seems that current reviews and discussions on ML applications in the sciences are overly focused on DL (Borowiec et al., 2022; Christin et al., 2019), which is due to the exaggerated enthusiasm in the tech industry that can tempt researchers in the life sciences to miss the opportunity and not consider "simpler" ML models. This bias has the potential to overshadow the benefits of other ML algorithms that may themselves offer unique advantages.

The rising popularity of DL requires a conscientious approach to scientific practise, leading us to question whether DL actually offers the most appropriate tools for solving a particular scientific problem. In the life sciences, the ultimate goal of ML revolves around understanding – deciphering the mechanisms of the biological system that generate the underlying data. However, this endeavour can be hindered when dealing with deep ANNs, which are "black boxes" due to their inherent complexity. Such models prevent us from understanding what the model has actually learnt.

There is a widespread misconception in the scientific community that there is a trade-off between accuracy and interpretability and that more complex models inevitably lead to higher accuracy. This leads to the assumption that complex 'black box' models are mandatory to achieve better predictive performance. Contrary to this assumption, simpler models often perform better, especially when the data is well structured and can be expressed in terms of naturally meaningful features (Rudin, 2019). Moreover, there is little evidence that DL outperforms classical ML algorithms on supervised learning tasks with limited structured (tabular) data. In reality, the additional flexibility of DL algorithms only bears fruit when the available data is both sufficiently large and complex (Pichler et al., 2020).

Given these observations, it is important to carefully consider whether the use of DL is actually necessary or even beneficial for a particular task. Simpler models can offer better interpretability, superior statistical power and lower computational costs, thus reducing the environmental footprint (Schwartz et al., 2020). This is particularly important in the context of certain life science sub-disciplines such as ecology, where the scarcity of data can make the use of large ANNs inappropriate.

When dealing with data consisting of a clear set of features structured in tabular form, simpler ML models such as decision trees (DT), Random Forest (RF) (Breiman, 2001), Support Vector Machines (SVM) (Cortes & Vapnik, 1995) or shallow Multilayer Perceptrons (MLP) and Neural Networks (Ripley, 2007) may outperform their more complex counterparts.

Certain tasks, such as automatic species identification, indeed benefit from the use of DL algorithms as they can better process spatial patterns (Ferreira et al., 2020; Lecun et al., 2015; Tabak et al., 2020). Nevertheless, traditional ML algorithms tend to deal more effectively with a smaller number of observations or structured (tabular) data (Arik & Pfister, 2021; Pichler et al., 2020).

It is therefore not surprising that, although the use of ML and DL methods in ecology and evolution research has increased considerably in recent years, classical ML methods are still predominant in practical applications (Pichler & Hartig, 2023), and it is projected that these classical methods will continue to hold significant importance in the future given the nature of the tasks in ecology and evolution (Pichler & Hartig, 2023)(Pichler and Hartig, 2023). One important reason for this is certainly their explanatory power.

1.3 Explainability: A Prerequisite for Scientific Inference

In this section we discuss topics related to explainability in ML. We mention some of the approaches and techniques and argue why they are relevant for applications of ML in the life sciences. First, however, we need to clarify some terms related to explainability that are used in this dissertation.

1.3.1 Explainability vs. Interpretability

There is no real consensus on what interpreatability or explainability is in ML (Molnar, 2022). Some authors speak of the unresolved interpretability problem, which arises from the fact that interpretability is a very subjective concept and therefore difficult to formalise (Rüping, 2006). The necessity of interpretability arises from the incompleteness of the problem formalisation (Doshi-Velez & Kim, 2017). Carvalho et al., 2019 explain that for certain problems it is not sufficient to obtain the prediction (*the what*). As they go on to explain, the model must also "explain how it arrived at the prediction (*the why*), because a correct prediction only partially solves the original problem" (Carvalho et al., 2019).

In this dissertation we use the two terms *explainable* and *interpretable* interchangeably, as is often the case (Molnar, 2022), which is not surprising considering that to interpret means to explain the meaning of something (Masís, 2023). Miller (2019) defines *interpretability* as the degree to which a human can understand the cause of a decision. Explainable or interpretable ML is then a "method and models that make the behaviour and predictions of ML systems understandable to humans" (Molnar, 2022).

Some authors, such as Masís (2023), try to distinguish between the two terms and define them as follows:

> **Interpretability** *is the extent to which humans, including non-subject-matter experts, can understand the cause and effect, and input and output, of a machine learning model.*
>
> **Explainability** *encompasses everything interpretability is. The difference is that it goes deeper on the transparency requirement than interpretability because it demands human-friendly explanations for a model's inner workings and the model training process, and not just model inference.*

We tend to use *explainable ML* to capture the broader meaning of explainability. Scientific and real-world applications of ML solutions, which are the focus of this dissertation, require transparency and trustworthiness for stakeholders, which goes beyond understanding the predictions of the model and thus beyond mere interpretability. Another reason is the alignment with the now well-established AI sub-discipline called explainable artificial intelligence (xAI), which we discuss in more detail below.

1.3.2 xAI Goals and Explainability Methods

xAI is concerned with research and practises that focus on the development of AI systems that are transparent and understandable through the application of explainability methods. We use the term xAI in a similar way to explainable ML, as both have similar goals, while the first term is broader.

As can be observed the motivations of xAI for a better understanding or explanation of ML systems are manifold. Arrieta et al. (2020) has produced an extensive literature review of papers on xAI, which showed that there is not much consensus on what ML explainability should compel. However, the review synthesised the most common goals of xAI: trustworthiness, causality, transferability, informativeness, confidence, fairness, accessibility, interactivity, and privacy awareness. As will be shown below, many of these goals are also of high priority in the application of ML in the life sciences.

xAI is primarily seen as an approach that addresses the black box nature of ML models by helping to understand how they make predictions (Ribeiro et al., 2016; Ryo et al., 2021). This is done using methods that provide insights about the input features' importance and an understanding of the global behaviour of the model, turning black box models into more translucent ones (Lucas, 2020). However, caution is required when interpreting the relationships inferred by ML, as optimising ML algorithms for predictive performance does not guarantee a correct understanding of the causal dependencies between the input features and the response (Pichler & Hartig, 2023). On the other hand, xAI can also be seen as an approach that encourages the use of less complex – interpretable – ML models when possible.

When discussing xAI approaches, it is therefore necessary to further distinguish between *intrinsic* and *post hoc* explainability methods. *Intrinsic* methods achieve explainability by limiting the complexity of the ML model by using ML models that are considered directly interpretable by humans due to their simple structure (e.g. DT, linear or logistic regression). In contrast, *post hoc* methods analyse the model by applying explainability methods after model training (e.g. permutation feature importance) (Carvalho et al., 2019; Molnar, 2022), making the behaviour and predictions of complex ML systems more understandable to humans. There are many different types of post-hoc explainability techniques, which can be categorised into the following types depending on the author's intention, the method used and the type of the data: text explanations, visual explanations, local explanations, explanations by example, explanations by simplification and feature relevance explanation techniques (Arrieta et al., 2020). See Figure 1.1 for an overview of the post-hoc explainability techniques presented by the same authors.

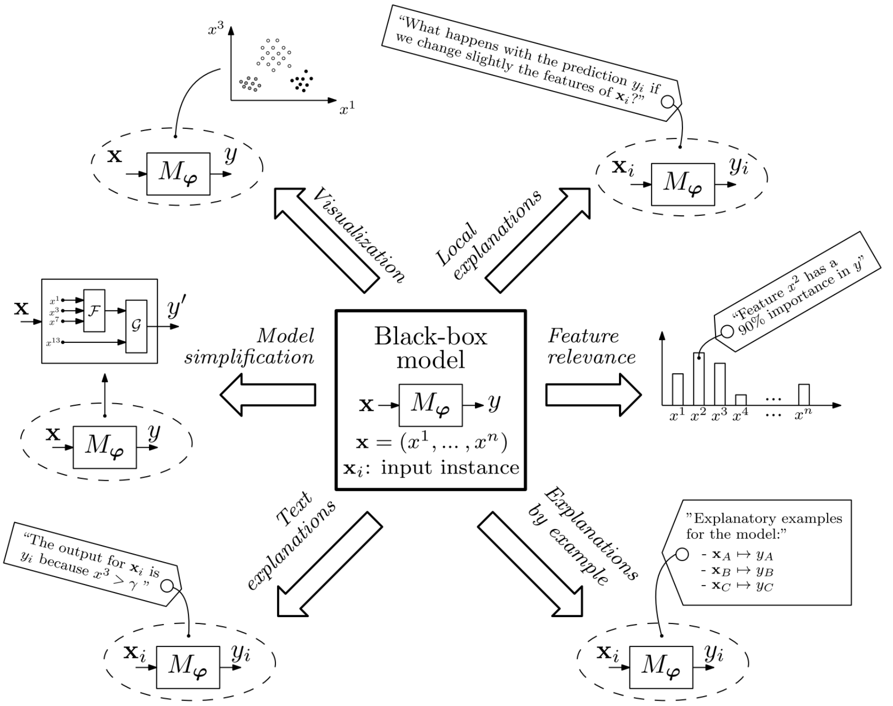


Figure 1.1: Conceptual diagram showing the different post-hoc explainability approaches available for a ML model (Arrieta et al., 2020).

Another useful division of xAI (Raghu & Schmidt, 2020) divides the broad suite of xAI techniques into two classes: The first is *feature attribution*, which provides insights about the importance of the input features. It can be used to provide feature importance for a single input and prediction. Examples of this are the feature ablation and perturbation methods, such as the best known SHAP based on the notion of Shapley value (Shapley, 2020) and LIME (Ribeiro et al., 2016). The second are *model inspection* and *representation analysis*, which explain the global behaviour of the model to understand the salient properties of the data, such as with probing and activating hidden neurons in an ANN, examples being network dissection (Bau et al., 2017), feature visualisation (Olah et al., 2017) and dimensionality reduction on NN hidden representations such as PCA, t-SNE (Van der Maaten & Hinton, 2008), UMAP (McInnes et al., 2018).

In this dissertation, intrinsically interpretable models and post-hoc explainability methods as well as feature attribution and model inspection techniques are applied in the context of life sciences. Among the explainability techniques in Figure 1.1, visualisation, feature relevance and model simplification are the most prominent. Further details can be found in the published scientific contributions below.

1.3.3 Explainable Machine Learning and Life Sciences

As shown in Section 1.2, ML offers compelling benefits for the life sciences, but ML models often prove difficult to interpret. This can lead to problems with trustworthiness, as potential biases can be obscured (Heil et al., 2021). In addition, the non-parametric nature of ML models implies complexity that, while good for capturing non-linear relationships, comes at the price of a higher number of parameters that often lack uncertainty estimates (Domingos, 2012). With the increasing use of ML in the life sciences, the importance of transparency in the areas affected by this scientific discipline, such as health, environment and food, is further increased. Predictions and research in the life sciences often carry significant weight for policy, decision-making and public health. Therefore, it is essential to carefully consider predictions for decision making, as they can have profound implications for the issues studied and trust between modellers, policy makers and other stakeholders (Lucas, 2020). Given these requirements and challenges, some researchers even recommend using interpretable models over explainable black boxes, especially for high- stakes decisions (Rudin, 2019). As Molnar (2022) points out, the easiest way to achieve interpretability is to only use algorithms that create interpretable models.

Nevertheless, with the popularity of powerful but opaque ML models such as RF and ANN in life sciences, xAI is becoming an indispensable approach in many subfields of life sciences such as ecology and evolution (Pichler & Hartig, 2023). Even when non-interpretable ML models are used to model complex non-linear problems, which are prevalent in fields such as ecology, stakeholders need to understand the model's decisions in order to trust it, which has led to a greater societal demand for xAI and trustworthy AI. When using ML to model harmful algal blooms (HABs) and their impact on aquaculture (e.g. (Brown et al., 2022; Karlson et al., 2021; McKenzie et al., 2021)), – which is one of the contributions of this dissertation – trust in ML-based early warning systems (EWS) is of paramount importance (Marzidovsek et al., 2024). If an AI system cannot provide a clear explanation for its decision-making process, this can diminish people's trust in the accuracy and reliability of outputs (Holzinger et al., 2023) on which human health depends. Besides requirements, the dissertation also identified causality, informativeness, accessibility and interactivity as important drivers for xAI approaches in the life sciences.

1.4 Motivation

The growing complexity of ecological data, coupled with the sophisticated nature of state-of-the-art ML methods, requires a collaborative approach between the ML and ecology communities (Tuia et al., 2022). High-level ML libraries have streamlined the ML pipeline (Caret, 2018; Pedregosa et al., 2012; Phillips & Dudík, 2008; Thuiller et al., 2016) and made it more accessible (Lucas, 2020). However, the integration of ecological domain knowledge into ML methods is essential to develop accurate models that are not only tools for prediction but also for understanding natural phenomena (Pichler & Hartig, 2023). This also applies to other sub-disciplines of the life sciences.

This work therefore sets out to collaborate with ecologists, marine and terrestrial biologists as well as experts from agricultural science to overcome some of the challenges and advance the use of ML in the selected sub-disciplines of life sciences. Some of the challenges identified in the selected area are the heterogeneous data types, complex data aggregation, missing values, class imbalances and other data preprocessing related issues. However, the most important aspect for the use of ML methods in life sciences is the need for greater transparency and a better understanding of the inner workings of the learnt models on the one hand and satisfactory predictive performance on the other.

As shown above, ML can accelerate and improve the quality of life sciences research, with important implications for society and its transition to more sustainable patterns of living. Insights into the dynamics of modelled ecosystems are critical to our food security and our understanding of the underlying factors associated with global challenges such as food chain disruption, biodiversity loss and climate change. The aim of the dissertation is to advance the use of ML, network science and xAI methods in selected sub-disciplines of life sciences to promote research that can provide answers and solutions to address the escalating impacts of climate change and human-induced stressors on biodiversity that can lead to ecosystem collapse.

1.5 Goals and Hypotheses

The goals of the dissertation correspond to the results of the published work and are as follows:

- G1: Develop models for direct prediction of the impact of harmful algal blooms (HABs) on shellfish aquaculture in the Slovenian coast. Interpretable models are used and xAI methods applied on the opaque models to increase the trustworthiness so that they can be used in real-world EWS for mussel contamination. In addition, the application of xAI methods can lead to a better understanding of marine ecosystem dynamics due to environmental and climatic pressures.

- G2: Design the necessary data pre-processing steps and integrate them into the ML pipelines to achieve better results when training on challenging real-world ecological data from different sources. This will help ecologists apply the pre-processing methods required for state-of-the-art computational modelling with heterogeneous, imbalanced, limited and noisy data common in the ecological domain.

- G3: Identify the barriers and incentives to the use of decision support systems (DSS) for integrated pest management (IPM), which is crucial for mitigating the negative impacts of agriculture on natural ecosystems and biodiversity. The problem is addressed using unsupervised ML and post-hoc explainability methods such as dimensionality reduction and clustering, visualisation explanations and explanations by simplification.

- G4: Create a state-of-the-art computer vision system for object detection and instance segmentation to automate the analysis of changes in phytoplankton community structures over time. Visual explanations are applied for greater thrustworthiness and confidence in the reliability of the ML based solution.

- G5: Develop extensible and easy-to-use software tool for network analysis and ML on graphs with explainability features that can help ecologists to better understand the topology and dynamics of ecological networks. This is pursued with various network analysis metrics, interactive network visualisation and representation learning on graphs. The aim is to make network analysis and state-of-the-art ML algorithms more accessible to ecologists. The tool was developed in collaboration with ecologists and evaluated through a case study that assesses the biodiversity-ecosystem functioning on a soil food web of one agricultural grassland.

The hypotheses investigated in this dissertation are:

- H1: The benefits of employing simpler and interpretable ML models for ecological modelling of HAB impacts on shellfish may outweigh potential limitations in predictive performance.

- H2: Top-performing, sophisticated ML algorithms are suitable for ecological modelling of HAB impacts on shellfish when coupled with explainability methods.

- H3: Unsupervised ML and explainability methods can reveal user attitudes towards utilising DSS in IPM.
- H4: Effective computer vision models for marine biology research can be developed via transfer learning on limited training data, and their trustworthiness increased with explainability methods.
- H5: Specially developed software integrating accessible network analysis, ML, and explainability methods can enhance understanding and knowledge discovery from ecological networks.

1.6 Methodology

The methodology of the dissertation consists of a literature review, defining gaps and setting research goals, data selection and preprocessing, network analysis, algorithm selection, training ML models, performance evaluation and comparison, model explanation, and evaluation of results by domain experts. The specific methods are described in the individual chapters of the dissertation sections, where the published work is presented.

1.7 Contributions

The dissertation presents the developed ML approaches and research work from marine biology, ecology and agricultural science, in which domain experts were involved throughout. The work demonstrates the application of supervised and unsupervised ML methods using tabular, time-series, graph structured and image data. Central considerations are the development and application of techniques that enable (1) effective ML on smaller, noisy and imbalanced datasets, as well as (2) better explainability and trustworthiness of ML models, both of which are critical for applications in the life sciences.

In addition, a need for visualisation and analysis of ecological networks and more accessible ML methods has been identified in the field of ecology. This gap has been filled by the development of a software tool focusing on the analytical requirements of ecological research.

The scientific contributions of the dissertation addresses pertinent problems of ML applications in life sciences, such as the need for customised and accessible network analysis, ML and xAI methods that provide more transparency and understanding of models. The scientific relevance is demonstrated through four publications in international peer-reviewed journals with impact factors and a publication in the proceedings of an international scientific conference. More detailed information about the underlying work and publications can be found below:

1. Marzidovšek, M., Francé, J., Podpečan, V., Vadnjal, S., Dolenc, J., & Mozetič, P. (2024). Explainable machine learning for predicting diarrhetic shellfish poisoning events in the Adriatic Sea using long-term monitoring data. *Harmful Algae*, 102728.

Training and evaluation of unique models capable of direct prediction of toxicity in shellfish aquaculture. Due to xAI methods, the models provide the necessary transparency so that the models, otherwise percieved as black box models, can be applied in real-world scenarios for early warning systems. The applied xAI methods also provide a better understanding of the dynamics of the marine ecosystem due to environmental and climatic pressures. Aggregation of a new dataset from the North Adriatic region spanning almost three decades and combining biological, chemical and environmental data has also been gathered.

This work and the publication contribute to the the dissertation goals G1 and G2.

2. Marinko, J., Ivanovska, A., Marzidovšek, M., Ramsden, M., & Debeljak, M. (2023). Incentives and barriers to adoption of decision support systems in integrated pest management among farmers and farm advisors in Europe. *International Journal of Pest Management*, 1-18.

Application of unsupervised learning methods and post-hoc explainability methods such as clustering, dimensionality reduction, visualisation explanations and explanations by simplification to promote IPM. To our knowledge, no one has yet attempted to apply unsupervised ML methods such as dimensionality reduction and clustering to such data to identify the barriers and incentives for using DSS for IPM, which is critical for mitigating the negative impacts of agriculture on natural ecosystems and biodiversity.

This work and the publication contribute to the dissertation goal G3.

3. Marzidovšek, M., Mozetič, P., Francé, J., & Podpečan, V. (2024). Computer Vision Techniques for Morphological Analysis and Identification of Two Pseudo-nitzschia Species. *Water*, 16(15), 2160.

Computer vision method to identify and quantify phytoplankton species, estimate cell size and calculate biovolume based on cell morphology. The approach involves transfer learning, where a generic pre-trained model for instance segmentation is fine-tuned with microscope images. Compared to manual analysis, this method processes more samples and provides a more ecologically relevant assessment with high confidence due to the visual explanations.

This work and the publication contribute to the dissertation goal G4.

4. Marzidovšek, M., Podpečan, V., Conti, E., Debeljak, M., & Mulder, C. (2022). BEFANA: A tool for biodiversity-ecosystem functioning assessment by network analysis. *Ecological Modelling*, 471, 110065.

It is a tool with unique analytical methods that was developed in collaboration with ecology experts and evaluated through application in a case study. It fills the gap as dedicated tools for ecological research that provide various network analysis metrics, interactive network visualisation and representation learning on graphs are not sufficiently accessible.

This work and the publication contribute to the dissertation goals G2 and G5.

1.8 Organisation of the Dissertation

The rest of the dissertation is organised as follows: Chapter 1 provides the necessary background on ML applications in the life sciences. It provides the general motivation as well as the goals and the five hypotheses that are investigated in this dissertation. The chapter concludes with a list of all contributions and publications in which the research results have been published.

The contributions to the dissertation are described in detail in chapters 2 to 5. Chapters 2 to 4 each contain a paper published in an international journal in which the contributions were originally presented, as well as an introduction describing the methodology used and the hypotheses addressed. Chapter 2 presents explainable ML for the prediction of shellfish toxicity. Chapter 3 presents unsupervised ML in the field of IPM. Chapter 4 presents the analysis of phytoplankton community structure using computer vision. Chapter 5 presents BEFANA: A tool for network analysis and ML on graphs in ecology. Finally, Chapter 6 concludes the dissertation and gives an overview of all contributions.

Chapter 2

Explainable Machine Learning in Marine Ecology

This chapter presents the first contributions to the dissertation that relate to the application of ML in ecology, especially in marine biology, through the analysis of long-term structured data.

Explainable ML techniques were applied to predict the toxicity of mussels in the Gulf of Trieste (Adriatic Sea) caused by harmful algal blooms (HABs). By analysing a newly created 28-year dataset containing records of toxic phytoplankton in mussel farming areas and toxin concentrations in mussels, we trained and evaluated the performance of ML models to accurately predict diarrhetic shellfish poisoning (DSP) events. xAI methods are applied with the aim of gaining insights into the model's decisions.

The development of the ML approach comprised the following main steps and methods. First, data from different sources were aggregated and matched to create a new consolidated dataset spanning almost three decades and combining biological, chemical and environmental data. After feature selection missing data were removed. The remaining missing data were interpolated using values within a specific time frame. This was followed by a descriptive analysis, dimensionality reduction and visualisation in 2D. The dataset was cleaned by removing samples close to the decision boundary using EditedNearestNeighbours, which was implemented in scikit-learn. Then the dataset was split into a training and a test set (70:30).

This was followed by the learning-oriented steps. The models were trained and evaluated for the direct prediction of shellfish toxicity. The following ML algorithms were used: SVM, decision tree, random forest, MLP. 5-fold cross-validation, hyperparameter optimisation with GridSearch was used to select the best models based on F1 score to balance the performance of both on recall and precision. Upsampling (SMOTE) and downsampling (RandomUnderSampler) have been integrated into the GridSearch pipeline to address the significant class imbalance.

Third, the evaluation and analysis steps followed. The best performing models (F1-score) from GridSearch were validated on the hold-out test set and the performance was measured with recall, precision and the F1-score. Finally, all algorithms were compared based on the performance metrics. The entire training and evaluation was repeated 100 times and the results were averaged to obtain more stable performance.

Finally, the decision tree model was visualised for better interpretability and the best performing black box models, random forest explained using post-hoc explainability methods permutation importance and SHAP. This provided model inspection of the global model behaviour while feature attribution based on SHAP explained the individual predictions of the models.

The implementation of the ML approach consisted of the following technologies and tools: Python within JupyterLab, the latest interface of the Jupyter project (Kluyver et al., 2016), the scikit-learn machine learning library (Pedregosa et al., 2012), pandas (McKinney et al., 2010) and for visualisations seaborn (Waskom, 2021) and Matplotlib (Hunter, 2007).

The results show the following. Firstly, ML models, especially RF, can satisfactorily predict DSP toxicity in mussels from Slovenian mussel farms. Secondly, both RF explainability methods and DT visualisations show that Dinophysis fortii and Dinophysis caudata, together with abiotic factors influencing salinity of coastal waters, have the greatest influence on the predictions. Thirdly, the right data pre-processing steps are crucial for overcoming the specific challenges of consolidated datasets from different sources and determine the training and performance of the model. Fourth, the insights about the model's behaviour gained through explainability methods make ML approaches suitable for EWS due to the increased trustworthiness. Such ML-based EWS can be a cost-effective means of implementing timely and appropriate mitigation measures in the aquaculture industry, such as trade bans, while improving management strategies to minimise health risks and social and economic damage.

This work directly addresses the following hypotheses defined in the introduction:

H1: The benefits of employing simpler and interpretable ML models for ecological modelling of HAB impacts on shellfish may outweigh potential limitations in predictive performance.

H2: Top-performing, sophisticated ML algorithms are suitable for ecological modelling of HAB impacts on shellfish when coupled with explainability methods.

Hypotheses H1 and H2 are fully confirmed by the results presented in this chapter.

The research work is presented in the following publication and is listed below:

Marzidovšek, M., Francé, J., Podpečan, V., Vadnjal, S., Dolenc, J., & Mozetič, P. (2024). Explainable machine learning for predicting diarrhetic shellfish poisoning events in the Adriatic Sea using long-term monitoring data. *Harmful Algae*, 102728.

Martin Marzidovšek contributed to this paper as follows: conceptualisation, data collection, methodology, software, formal analysis, writing the original draft, revising and editing, visualisation.

Harmful Algae 139 (2024) 102728

Contents lists available at ScienceDirect

Harmful Algae

journal homepage: www.elsevier.com/locate/hal

Explainable machine learning for predicting diarrhetic shellfish poisoning events in the Adriatic Sea using long-term monitoring data

Martin Marzidovšek [a,d,*], Janja Francé [b], Vid Podpečan [a], Stanka Vadnjal [c], Jožica Dolenc [c], Patricija Mozetič [b]

[a] *Jožef Stefan Institute, Department of Knowledge Technologies, Slovenia*
[b] *National Institute of Biology, Marine Biology Station Piran, Slovenia*
[c] *University of Ljubljana, Veterinary Faculty, National Veterinary Institute, Slovenia*
[d] *Jožef Stefan International Postgraduate School, Slovenia*

ARTICLE INFO

Edited by Dr. Satoshi Nagai

Keywords:
Harmful algal blooms
DSP toxins
Machine learning
Explainable artificial intelligence
Aquaculture
Marine ecology
Adriatic Sea

ABSTRACT

In this study, explainable machine learning techniques are applied to predict the toxicity of mussels in the Gulf of Trieste (Adriatic Sea) caused by harmful algal blooms. By analysing a newly created 28-year dataset containing records of toxic phytoplankton in mussel farming areas and diarrhetic shellfish toxins in mussels (*Mytilus galloprovincialis*), we train and evaluate the performance of machine learning (ML) models to accurately predict diarrhetic shellfish poisoning (DSP) events. Based on the F1 score, the random forest model provided the best prediction of toxicity results at which the harvesting of mussels is stopped according to EU regulations. Explainability methods such as permutation importance and Shapley Additive Explanations (SHAP) identified key species (*Dinophysis fortii* and *D. caudata*) and environmental factors (salinity, river discharge and precipitation) as the best predictors of DSP toxins above regulatory limits. These findings are important for improving early warning systems, which until now were based solely on empirically defined alert abundances of DSP species. They provide experts, aquaculture practitioners, and authorities with additional information to make informed risk management decisions.

1. Introduction

Over the last 30 years, shellfish aquaculture has grown steadily and accounted for 25% of global marine and coastal aquaculture production in 2020, which is crucial for food security (FAO-IOC-IAEA 2023). However, for certain species such as mussels, production in the European Union has declined due to challenges such as low profitability, limited progress in farming, and environmental threats such as harmful algal blooms (HABs), diseases, predators, unfavourable weather conditions and pollution (Avdelas et al., 2021). These global challenges are exacerbated by climate change, which call for an update of existing common practices and guidelines for their detection and appropriate management.

The term HABs refers to either non-toxic microalgae that reach high biomass and cause water discoloration, anoxia and mucilage formation that negatively impact the environment and human activities, or toxic species that threaten the safety of seafood and marine life (reviewed in Sagarminaga et al. 2023). The toxin-producing species that can cause food poisoning with neurological or gastrointestinal symptoms in humans are responsible for about 48% of the documented global HAB events recorded in the Harmful Algal Event Database (HAEDAT; https://haedat.iode.org/). In these events, toxins accumulate in wild and cultured shellfish and fish, severely impacting food safety and security (Hallegraeff et al. 2021).

In the Mediterranean region, around three-quarters of recorded toxic events involve diarrhetic shellfish poisoning (DSP) caused by dinoflagellates like *Dinophysis, Phalacroma* and *Prorocentrum* (Zingone et al. 2021). Paralytic shellfish poisoning (PSP), which is attributed to dinoflagellate species of the genus *Alexandrium* and *Gymnodinium catenatum*, and amnesic shellfish poisoning (ASP), which is attributed to several toxin-producing diatom species of the genus *Pseudo-nitzschia*, make up the rest of the toxic events (Zingone et al. 2021). Also in the Adriatic Sea, DSP toxins of the okadaic acid group together with other lipophilic toxins such as yessotoxins and pectenotoxins, are most frequently detected in shellfish and exceed legal limits (Accoroni et al. 2024; Nincevic Gladan et al. 2011), while ASP and PSP toxins represent

---

* Corresponding author at: Jožef Stefan Institute, Jamova 39, Ljubljana, 1000 Slovenia.
E-mail address: martin.marzidovsek@ijs.si (M. Marzidovšek).

https://doi.org/10.1016/j.hal.2024.102728
Received 9 May 2024; Received in revised form 17 September 2024; Accepted 18 September 2024

only a low risk for the time being (Ciminiello et al. 2005; Ujević et al. 2012).

In our area of interest, the Gulf of Trieste (northern Adriatic), farmed Mediterranean mussels (*Mytilus galloprovincialis*) are the main source of potential human poisoning. The monitoring programme for toxins in mussels and toxic phytoplankton in seawater has been ongoing since 1994. Patterns suggest a higher risk of exceeding regulatory limits from September to November consistent with the presence of DSP-producing dinoflagellates (Henigman et al. 2024). The predominant DSP-producing species, *Dinophysis caudata, D. fortii, D. sacculus, D. tripos*, and *Phalacroma rotundatum*, show a typical pattern of occurrence that has not changed significantly over the last 20 years and corresponds well with environmental characteristics such as a thermohaline stratified water column and higher temperatures (France and Mozetic 2006; Henigman et al. 2024). Despite robust monitoring systems aligned with EU regulations, complete protection against contaminated seafood cannot be guaranteed. The harvesting of mussels is suspended only when toxicity limits are exceeded, leading to periods of uncertainty due to the time required for analyses. In some cases, even changes in toxic plankton levels can result in precautionary closures until toxicity results are available. Furthermore, the usual pattern of toxic species, which is determined by the prevailing oceanographic and hydrological conditions, and the associated occurrence of toxins can also be disrupted by the unpredictable temporal variability of HABs, as has been observed in the Adriatic Sea and may be related to extreme weather events such as floods and droughts (Zingone et al. 2021). These situations illustrate that untimely or incorrect measures, such as overly strict precautionary closures in the event of negative toxicity results, can pose risks to human health or cause undue economic damage to shellfish farmers.

Given the negative economic impact, solutions are urgently needed to help the shellfish industry anticipate and adapt to HAB events that may lead to contamination and to support authorities in managing and mitigating risks such as the closure of shellfish production areas.

A recent comprehensive overview of early warning systems (EWS) for HABs (FAO-IOC-IAEA, 2023), pointed out that there is no single approach. The choice of observational technologies depends on the target organisms, types of HABs, and regions with specific environmental conditions as well as data resolution. For toxic pelagic HABs, where the causative organisms are present in low abundance—such as the case of *Dinophysis* species in the Adriatic Sea—various statistical and rule-based modelling, along with more complex machine learning and deep learning algorithms, demonstrate promising results in predicting toxicity in shellfish (Grasso et al. 2019).

Access to vast amounts of physico-chemical and biological data from multiple sources due to technological advances in meteorology and oceanography has led to the increasing use of machine learning (ML) techniques to predict HABs (e.g. Guallar et al. 2016; Derot, Yajima, and Jacquet 2020; Y. Park et al. 2021). ML can handle large and heterogeneous datasets and are powerful enough to model highly dynamic and nonlinear natural systems, making them suitable for accurately reproducing phytoplankton dynamics (Shimoda and Arhonditsis 2016), even when the data are noisy and the underlying relationships are not fully understood (Muttil and Chau 2006). However, a recent review of ML forecasting tools (Cruz et al. 2021) found that, in contrast to forecasting HAB species occurrence, very little has been done to foresee the toxicity of mussels (Bouquet et al. 2022; Grasso et al. 2019). This more difficult task is addressed in our study.

One limitation of ML techniques is that they often do not provide insights into the causal mechanisms of HABs (Recknagel, Orr, and Cao 2014). Simpler ML algorithms such as decision trees (DT) and linear regression are easier to interpret but cannot provide the required predictive power. Cruz et al. (2021) found that model complexity has increased in recent years, at the expense of explainability. However, when modelling scenarios that involve risks to human health and ecosystem disruption, as is the case with HABs, it is crucial not only to assess the reliability of predictions, but also to understand the drivers behind the model's decisions. For these reasons, this study focuses on the training of interpretable ML models and the application of explainability methods that can provide insights into the behaviour of the model. It thus fits into the paradigm of explainable artificial intelligence (XAI), which has recently gained prominence as it promises to mitigate the drawbacks of so-called "black-box" models.

Our study represents one of the first attempts to test the performance of ML models for the short-term prediction of DSP toxicity of mussels in the Mediterranean Sea, where DSP events are the major concern for seafood safety. A nearly 30-year dataset of phytoplankton species and toxin concentrations in bivalves was created along with key environmental data to train ML models, and both the dataset and the full code are openly available to stimulate further research (https://github.com/MartinMarzi/HABTox-predictor). The obtained ML models were also evaluated and interpreted based on real, long-term monitoring data.

The objectives of the study were: (1) to develop data preprocessing pipelines suitable for the specific requirements of the dataset; (2) to train selected ML models for direct prediction of DSP toxicity events and evaluate their potential for use in EWS; (3) to explain the obtained ML models and their predictions for improved model trustworthiness.

2. Material and methods

2.1. Site description

The Slovenian Sea is situated in the southeastern region of the Gulf of Trieste (GoT), marking the shallowest (average depth around 20 m) and northernmost part of the Adriatic Sea. The physical, chemical and biological characteristics of this semi-enclosed gulf (600 km2 in area) are influenced by local meteorological and hydrodynamic conditions. The latter are in turn determined by inflows from rivers, winds, tides and water exchange with the Adriatic Sea (Falcieri et al. 2016). In the surface layer, the currents are mainly driven by the wind, while in deeper layers the circulation pattern is predominantly cyclonic (Stravisi 1983). The residual currents are about 1–3 cm s–1, but total currents of >30 cm s–1 can occur in the upper layer (Mosetti and Purga 1990).

The Soča River, which flows into the Gulf on the shallower northern coast with an average flow of 123 m3 s–1, represents the largest freshwater input and the main source of land-based nutrients (Comici and Bussani 2007), while the rivers along the Slovenian coast discharge <10% of the freshwater into the Gulf and provide an even smaller nutrient contribution (Cozzi et al. 2012). During most of the year, the Soča River plume often spreads over a large area of the Gulf (Malačič and Petelin, 2001), resulting in a residence time of up to 23 days (Cantoni et al. 2003). Increased freshwater inflow in spring together with atmospheric warming leads to the formation of a pycnocline, which intensifies in summer, while the water column is mixed in winter (Malačič 1991). This dynamic interplay of physical and chemical properties is reflected in the seasonal and interannual variability of the phytoplankton community (Mozetič et al. 2012). The increasing warming of seawater, alternating droughts and floods, combined with human activities in the watershed, have led to an imbalance of nutrients in coastal waters and a general decline in phytoplankton biomass not only in GoT but throughout the northern Adriatic (Brush et al. 2020).

Moreover, the Slovenian coastal sea faces substantial anthropogenic pressures. The area is highly urbanised, experiencing intensive land-based and nautical tourism, hosting an international cargo port, and witnessing an expanding aquaculture industry. In particular, mussel farming is practised in three protected bays on the Slovenian coast - Debeli rtič, Strunjan and Seča (Fig. 1) with annual mussel production up to 700 tonnes.[e] Surveillance of the seafood safety began in the late 1980s on an irregular basis, while the official monitoring programmes for toxic

---

[e] Statistical Office of the Republic of Slovenia: https://www.stat.si/statweb/en.

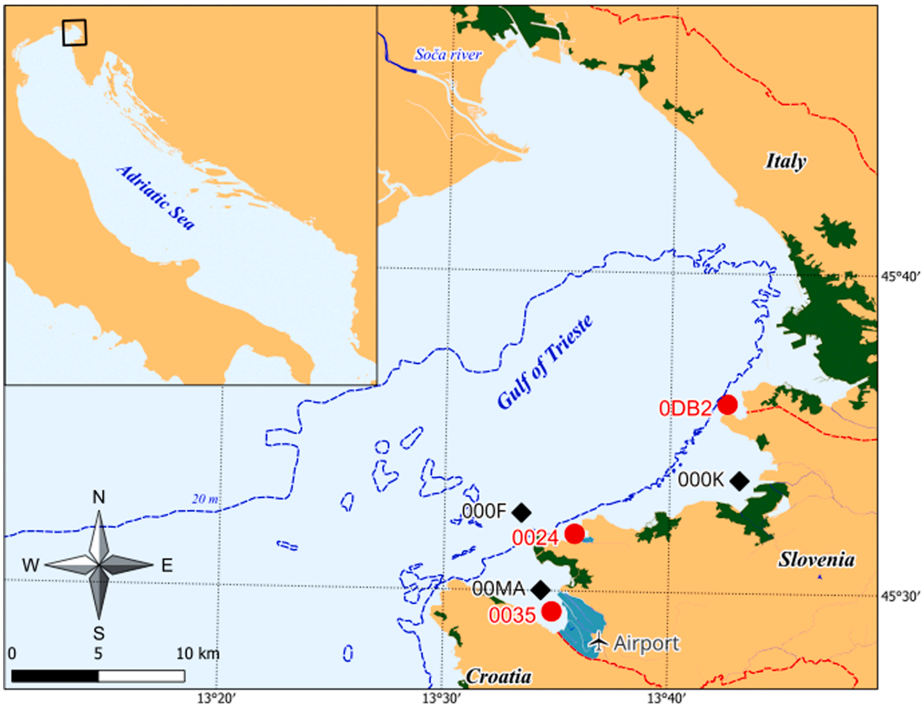


Fig. 1. Map of the GoT (Adriatic Sea) and location of the three mussel farming sites with the respective sampling stations on the Slovenian coast: Debeli rtič (0DB2), Strunjan (0024) and Seča (0035). Physical and chemical parameters of seawater were collected at sampling stations 000 K, 000F and 00MA as part of ecological monitoring in accordance with the Water Framework Directive 2000/60/EC. The airplane symbol marks the location of the Portorož airport meteorological station.

phytoplankton and toxicity in bivalve molluscs were introduced in 1994.

2.2. Data acquisition

The data, which covers the period from 1994 to 2021 comes from different monitoring programmes carried out by different institutions and with different sampling frequencies. The subsequent creation of a consolidated dataset for ML analysis required careful matching of all types of data to retain the largest possible number of instances (see section 2.3).

Toxic phytoplankton and DSP toxins. The seawater samples for the phytoplankton analysis and the mussels were collected at the three sampling stations 0035, 0024 and 0DB2 (depth of 12 to 18 m), each located in a mussel farming area (Fig. 1). For phytoplankton, the sampling period within the year, frequency and sampling method (Niskin bottles at discrete depths, vertical haul with plankton net) varied until 2008, while the program followed a consistent scheme thereafter (complete annual survey, weekly to monthly frequency depending on the season, integrated water sample with a PVC hose). These differences were taken into account when preparing the data for the ML models. The identification and enumeration (cells l^{-1}) of the HAB species was performed under the inverted microscope according to the method of Utermöhl (1958). Since we applied ML models to predict DSP events, which are the main problem in the area, we narrowed the selection of data to the five most abundant and recurrent DSP-producing species (*Dinophysis caudata, D. fortii, D. sacculus, D. tripos, Phalacroma rotundatum*), together with the abundance of all DSP-producing species found in seawater samples (*Dinophysis* spp., *Phalacroma* spp. and *Prorocentrum lima*).

A similar approach was used for toxins in mussels. Initially, the sampling program varied considerably from year to year and was often adapted to the results of the phytoplankton analyses, with less regular and frequent sampling in winter during the first decades of observation. From 2011, the scheme became uniform and, as with phytoplankton, covered the whole year, with more frequent sampling in the months of higher toxicity risk (Apr-Nov). Until 2014, the standard mouse bioassay method (Yasumoto, Oshima, and Yamaguchi 1978)) was used in accordance with Directive 91/492/CEE (Council of the European Communities 1991) for the determination of DSP toxins, i.e. okadaic acid and its derivatives, the dinophysistoxins, which are extracted together with other lipophilic toxins such as yessotoxins, pectenotoxins and azaspiracids. Acetone shellfish extract was evaporated to dryness and residues dissolved in 1% Tween 60. This solution was injected intraperitoneally into three mice (Albino Swiss, 18 g-22 g body weight) and the survival time was monitored 24 h. The sample was positive if two or all three mice died. The mouse bioassay, having various disadvantages, including ethical concerns, and low sensitivity and specificity with only two possible test results - positive or negative, was in 2014 replaced by the more sensitive and specific method of liquid chromatography-mass spectrometry (LC-MS/MS). Comparison study between mouse bioassay and chemical methods show good correlation of results (Prassopoulou et al. 2009). The LC-MS/MS method allows chromatographic separation of the individual groups of lipophilic toxins (for details see (Henigman et al. 2024)). In Slovenia, harvesting of bivalve molluscs is banned if the regulatory limit of 176 μg kg^{-1} for DSP toxins (160 μg OA equivalents kg^{-1} according to Regulation (EC) No 853/2004 (European Council 2004) plus measurement uncertainty of the national reference laboratory) is exceeded. All toxin analyses were performed by the National reference laboratory.

Physical and chemical properties of seawater. In addition to the two datasets described above, seawater temperature and salinity were also taken into account as they describe the site-specific environmental conditions. Monthly sea surface temperature (SST) and salinity data were obtained from the same sampling stations as for toxic phytoplankton (0DB2, 0024, 0035) or, in case of missing data, from the nearest sampling stations (marked with black diamonds in Fig. 1) included in the ecological monitoring according to the Water Framework Directive (WFD) 2000/60/EC. The mussel farming stations and the WFD stations had a similar depth range and were between 0.7 NM and 2.3 NM apart. Temperature and salinity were measured with a fine CTD probe (Sea & Sun Technology GmbH).

Meteorological and hydrological data. Meteorological observations such as daily average air temperature and wind speed, daily solar irradiance and daily precipitation at the main coastal meteorological station (Portorož Airport) were obtained from the Slovenian Environment Agency,[f] which also provided the daily averaged flow rates of the main freshwater source, the Soča River.

2.3. *Data preprocessing*

The performance of an ML model depends largely on the quantity and quality of the data used to train it. The right choice of input variables affects the predictive performance of the model, as does an imbalance between classes (Menardi and Torelli 2014), which can lead to biassed learning because the model learns to predict the majority class better. Our goal was to collect all relevant input variables from the different data sources (e.g. Cruz et al. 2021; Patrício et al. 2022), while obtaining a sufficient amount of data for ML training after preprocessing the missing values. We applied extensive data matching, aggregation and interpolation methods to obtain a consolidated dataset with 14 independent variables and DSP toxicity as the target variable, which were then used for training (Table 1).

The abundances of five phytoplankton species and the aggregated variable DSP-tot – the sum of the abundances of all DSP-producing species – were included. As sampling techniques changed during the monitoring program, transformation and validation were performed. First, the abundances from the net samples were multiplied by 100 to obtain the same order of magnitude as with other sampling techniques. Then we tested the differences between the relative frequency distributions of species abundances obtained with different sampling techniques with the Kolmogorov-Smirnov test. The test results showed that distributions of the five most abundant DSP species do not statistically significantly differ ($p > 0.05$) except for *D. fortii* between Niskin and hose samples where the distributions were slightly different ($p = 0.05$).

When multiple samples were from the same day and location but from different depths, the samples with the highest abundance of DSP-tot were selected.

The toxicity results also had to be standardised to take account of the different methods. In accordance with the legal threshold for DSP toxins (176 $\mu g\ kg^{-1}$), the concentrations of the chemical analyses (2014–2021) were mapped to binary positive or negative test results to match the results of the bioassays (1994–2013). The total number of toxicity tests received (binary target variable) was 1132, of which 996 (88%) were negative and 136 (12%) were positive.

Because the phytoplankton and toxicity datasets had different sampling frequencies and temporal resolutions, a matching toxicity test was assigned to a phytoplankton observation based on a time window, separately for the three sampling stations. A Python script selected the first possible toxin result based on the timestamp of a given phytoplankton observation, but only if it was within a time window of 30 days after the phytoplankton observation. This maximum acceptable time window was chosen because the frequency of sampling for toxic phytoplankton was at least once per month. Analysis showed that the average time window in the obtained consolidated dataset was 6 days and 75% of all examples had a time window of less than or equal to 9 days.

Before we could proceed with the aggregation of SST and salinity, the problem of missing data had to be solved. When SST and salinity data were not available from mussel farming stations, the closest possible values within a time window of 30 days prior the phytoplankton observation were taken from the nearest WFD monitoring stations. For the meteorological data, daily solar radiation, precipitation, air temperature and wind speed were added by averaging or summing (depending on the nature of the variable) for a time window of 20 days prior to the phytoplankton observation. Finally, the daily flow of the Soča River, being the good proxy for freshwater nutrient inputs in the GoT (Cozzi and Giani 2011) was summed over a period of 30 days prior to the phytoplankton observation. This time window was longer than that of the meteorological data, as the residence time of the Soča River is longer than 20 days, especially in periods of calm weather (Cantoni et al. 2003), while that of nutrients is even longer (Cozzi et al. 2012).

The dates in the consolidated dataset were converted to months to allow the model to better capture the annual cyclical patterns of phytoplankton growth. Through these steps, the consolidated dataset retained all 1452 instances from phytoplankton monitoring and, importantly, all toxicity results (1132). In a further preprocessing step, instances with missing values (44 in total) were removed, as some ML algorithms cannot process data with missing values.

We used Uniform Manifold Approximation and Projection (UMAP) (McInnes, Healy, and Melville 2018) for dimensionality reduction, which allowed to visualise the data in two dimensions to gain insight into how the data are partitioned with respect to the target variable. This method projected the consolidated dataset into a lower-dimensional feature space, attempting to preserve as much variance or structure as possible. The positive and negative examples were quite mixed throughout the data space (Fig. 2, left), making it more difficult for a model to learn a clear decision boundary between the data. Since the dataset was heavily imbalanced in favour of the negative class, we used the Edited Nearest Neighbours from the imbalanced-learn Python toolbox (Lemaître, Nogueira, and Aridas 2017) to remove instances from this class whenever they were close to instances with positive class values. This removed 211 conflicting examples with negative test results (Fig. 2, right). Our investigation confirmed that the removal of those examples improved the performance of ML algorithms.

The resulting dataset comprised a total of 877 instances (target class distribution: 745 negative, 132 positive). The dataset was then randomly split into a training and a test dataset, with a 70/30 split and stratified sampling to maintain the proportion of negative/positive examples in both datasets.

2.4. *Training machine learning models*

ML algorithms were used to train models for the prediction of DSP events in two binary classes - positive (1) or negative (0). One factor in the selection of ML algorithms and in determining the hyperparameters was the relatively small number of instances, especially of examples with positive DSP. In addition, the goal was to obtain ML models that could be directly explained and to use explainability methods for more opaque models. This influenced the selection of algorithms to tree-based algorithms such as decision tree (DT) and random forest (RF), support vector machines (SVM) and shallow artificial neural networks (ANN). These ML algorithms are widely used for ecological modelling and also for the prediction of HABs and DSP events (Grasso et al. 2019; Liu et al. 2022; Kim et al. 2021).

DT is a simple, interpretable model that splits data into branches to make predictions based on feature values. It is particularly useful in scenarios where interpretability is crucial and the problem is not overly complex as it builds decision rules that can be easily understood by aquaculture operators and other stakeholders. RF is an ensemble

---

[f] www.arso.gov.si

Table 1

Input variables for the training set with descriptive statistics over 28 years (1994 - 2021). DSP toxins were determined by mouse bioassay until 2013, while from 2014 they were chemically quantified (LC-MS/MS method). Concentrations above or below the legal threshold for DSP toxins were converted into positive or negative test results.

Variable	Description	unit	Min	Max	Mean	Median
month						
DSP-tot		cells l^{-1}	0	7634	96	30
Dinophysis caudata		cells l^{-1}	0	1309	23	0
Dinophysis fortii		cells l^{-1}	0	4624	23	0
Dinophysis sacculus		cells l^{-1}	0	4639	26	0
Dinophysis tripos		cells l^{-1}	0	1139	4	0
Phalacroma rotundatum		cells l^{-1}	0	393	13	5
SST		°C	6.23	28.87	17.41	17.28
salinity			24.13	38.66	36.54	36.99
air temperature	average air temperature in 20 days prior to phytoplankton sampling	°C	−1	27	17	18
wind speed	average wind speed in 20 days prior to phytoplankton sampling	m s^{-1}	1	5	3	3
precipitation	daily precipitation summed over 20 days prior to phytoplankton sampling	mm	0	268	59	45
solar irradiance	daily solar irradiance summed over 20 days prior to phytoplankton sampling	h	27	278	164	168
river flow	daily flow rate of the Soča River summed over 30 days prior to phytoplankton sampling	m^3 s^{-1}	594	16,040	3301	2580
DSP toxins	positive or negative result of toxicity test					

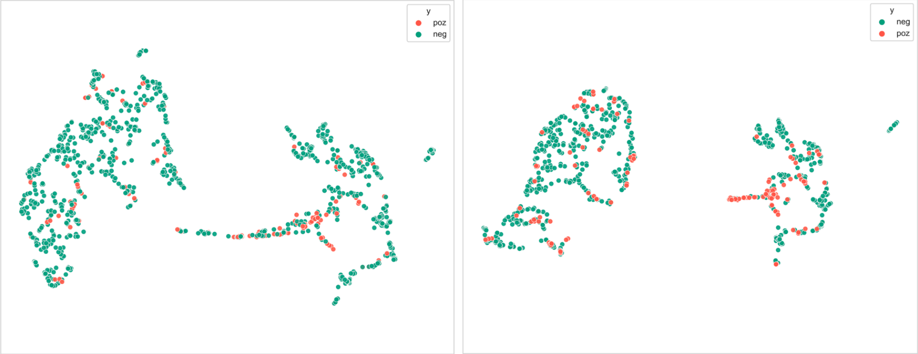


Fig. 2. UMAP projection of data in two dimensions before (left) and after (right) removing instances that are closely positioned in data space and have opposite class values. The data points labelled as "poz" are for positive toxicity examples and "neg" for negative toxicity examples.

learning method that combines multiple decision trees to improve predictive accuracy and control overfitting. It is widely used in ecological modelling because it can capture more complex nonlinear relationships and provides some explainability with built-in methods (feature ranking). SVM is a supervised learning model that finds the optimal hyperplane to separate different classes in the feature space. It is particularly effective in dealing with high-dimensional data and complex relationships between variables when the data have a clear margin of separation. An ANN is a computational model inspired by the human brain, consisting of interconnected layers of neurons that learn to make predictions from data. In the last decade, ANNs have become popular, especially deep learning methods, as they can model very complex nonlinear phenomena with potentially high predictive performance. However, they require a large amount of training data, are prone to overfitting and are considered black boxes as the decisions of the models are opaque (Recknagel 1997). Due to the limited amount of training data, a shallow feedforward neural network, the multilayer perceptron (MLP), was used. This algorithm required further preprocessing of the data as it works better when the data is standardised, for which z-score normalisation was used. The month variable was also removed.

As the training set was significantly imbalanced, we addressed this issue by undersampling and data augmentation. As Kim et al. (2021) have shown, synthetic data can improve the ML models' detection of HABs. Our method integrates under- and oversampling into the modelling pipeline, which systematically explores and optimises hyperparameters with a grid search over a predefined range of values to determine the most effective settings for data augmentation and ML algorithms. Criteria for splits, class weight adjustments and structural parameters such as tree depth and layer complexity were taken into account. This careful process ensured robustness against overfitting, especially given the constraints imposed by the size of the dataset.

In the first step, the training data (validation data) was oversampled using the Synthetic Minority Over-sampling Technique (SMOTE) to increase the number of instances with positive toxicity. The next step in the pipeline was to undersample the majority class (negative toxicity) of the training data to further reduce the class imbalance. In the third step, the ML models were trained with stratified 5-fold cross-validation on the over- and undersampled training set with the optimised parameter values. Several model performance metrics were calculated: (1) Precision: The ratio of true positive (TP) results to the total predicted positives, which measures the accuracy of positive predictions; (2) Recall: The ratio of TP results to the total actual positives, which measures the model's ability to capture all positive instances; (3) F1 Score: The harmonic mean of precision and recall, which provides a single metric that balances both concerns. Because this study had a significantly unbalanced dataset (with more negative toxicity outcomes than positive) the F1 score was advantageous because it balances precision and recall, providing a more meaningful measure of the model's performance on the minority class than accuracy alone. The optimal algorithm parameter configuration was therefore determined based on the highest mean F1 score of the 5-fold cross-validation from each grid search combination.

2.5. Evaluation of models

To assess the generalisability, the evaluation of the constructed ML models was performed on unseen examples and on data distributions they would encounter in real-world applications. Based on the performance of the obtained models, the highest-ranking model from grid search was then selected for each of the ML algorithms and subsequently evaluated on the test set that has been left out of the training pipeline. Finally, the training and evaluation on the test set was repeated 100 times, and the results of the selected performance metrics (precision, recall, F1-score) were averaged to obtain a more statistically reliable estimate of each model's performance. The most relevant metric for the task at hand was recall, as it indicates the proportion of predicted positive class instances relative to all TP instances in the test set. However, to ensure that the classifier does not predict every instance as positive and thus make too many false positive (FP) predictions, we also needed to control precision. Therefore, the F1 score was used in the parameter optimisation as it balances both recall and precision. Finally, the averaged performances for all three performance metrics were compared for all four ML algorithms.

2.6. Explaining ML models

This study places the same emphasis on the predictive power as on the explainability of the model. By applying XAI principles, the aim is to better understand the inner workings of models by identifying the most informative relationships between the input variables and the predicted target variable. XAI calls for the use of interpretable ML models whenever possible and the application of explainability techniques for opaque ML models: Shapley Additive Explanations (SHAP) (Lundberg and Lee 2017), Local Interpretable Model-Agnostic Explanations (Lime) (Ribeiro, Singh, and Guestrin 2016), permutation importance, various feature ranking methods, and others.

The DT used in the study could be inspected directly by visualising the tree structure of the model. For the more complex models – RF, SVM and ANN – selected explainability methods were used to gain insights into their behaviour. The importance of the variables, i.e. the feature importance, was determined using two selected model-agnostic methods that can be applied to different model types. Permutation feature importance from the scikit-learn library was used to obtain a ranking of how strongly each of the variables influences the RF model. This permutation-based method of variable importance reflects the decrease in a model's performance score when a single variable value is randomly shuffled (Breiman 2001). Shuffling removes the relationship between the independent and the target variable, leading to a decrease in model performance and thus showing how much the model depends on the particular variable.

The SHAP method was used to gain a better understanding of the contribution of the individual variables to the models' output. For RF, the SHAP TreeExplainer, an algorithm specifically for tree ensemble methods, was applied. This method allows a general model interpretation as it represents the behaviour of the model over the entire data set on which it was trained. In addition, it also allows us to inspect individual model's predictions which were also implemented for this study, as they are relevant for real-world deployment.

2.7. Implementation

The data preprocessing, the statistical analysis, the parameter optimisation and the model construction, evaluation and interpretation were carried out using the Python programming language. In the study, it was used in combination with JupyterLab, an interactive development environment for computational notebooks (i.e. Jupyter notebooks), and pandas (McKinney and Others 2010). ML algorithm implementations from scikit-learn (Pedregosa et al. 2011) and other specialised tools such as UMAP (McInnes, Healy, and Melville 2018), Imbalanced-learn (Lemaître, Nogueira, and Aridas 2017) and SHAP (Lundberg and Lee 2017) were used. The complete code and data are available in an online repository (https://github.com/MartinMarzi/HABTox-predictor).

3. Results

3.1. Analysis of long-term phytoplankton and DSP toxins monitoring

The basic oceanographic features in the Slovenian part of the GoT display notable annual fluctuations in seawater temperature and salinity, as shown in Fig. 3. From 1994 to 2021, SST fluctuated between 6.23 °C in February and 28.87 °C in July, while surface salinity fluctuated between 24.13 and 38.66 (Table 1), with the two extremes occurring in February. The lowest monthly salinities (shown as triangles in Fig. 3) were measured in the winter months (January-February) and in November, suggesting that less saline water is present in the surface layer during these periods. However, on average, the winter months had high salinity, while the late spring/summer months (May-July) and November had below average salinity. The highest temperature peaks for each month (also indicated by triangles in Fig. 3) were mostly observed in the last decade. The largest deviations from the average climatology were observed in spring/summer (May-July).

The annual distribution of DSP-tot (as 90th percentile), averaged over the 28-year time series, shows that the period from May to December is the most likely time for DSP toxicity events in the GoT (Fig. 4). Two peaks in abundance, reaching values of up to 7600 cells l^{-1} (Table 1), were observed in June/July and from September to November, corresponding to the distribution patterns of the five main species. The "blooms" of *Dinophysis sacculus* and *D. caudata* were typical of the early summer months, while *D. fortii* was responsible for the fall peak. During the rest of the year, the DSP-producing species were almost absent in seawater.

The DSP toxin analyses were intensified from May to November to follow the dynamics of the toxic species. Monthly testing varied greatly, ranging from about 20 tests in winter to about 170 tests in fall from 1994 to 2021 (Fig. 5). The number of positive tests increased significantly in the months with increased sampling, especially in September and October, where absolute values peaked at 31 and 24 tests, accounting for 23% and 18% of all positive tests, respectively. The incidence of DSP

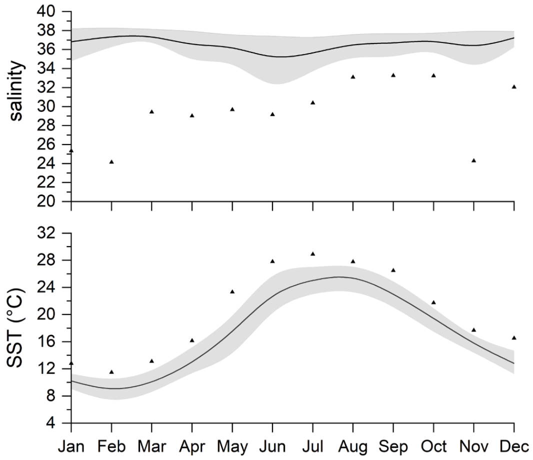


Fig. 3. Annual distribution of mean salinity and seawater temperature (SST) in the surface layer of Slovenian coastal waters, period 1994–2021. The grey band around the mean (line) is the 10–90 percentile. The triangles indicate the lowest salinity and the highest temperature for a given month measured at any point during the 28-year time series.

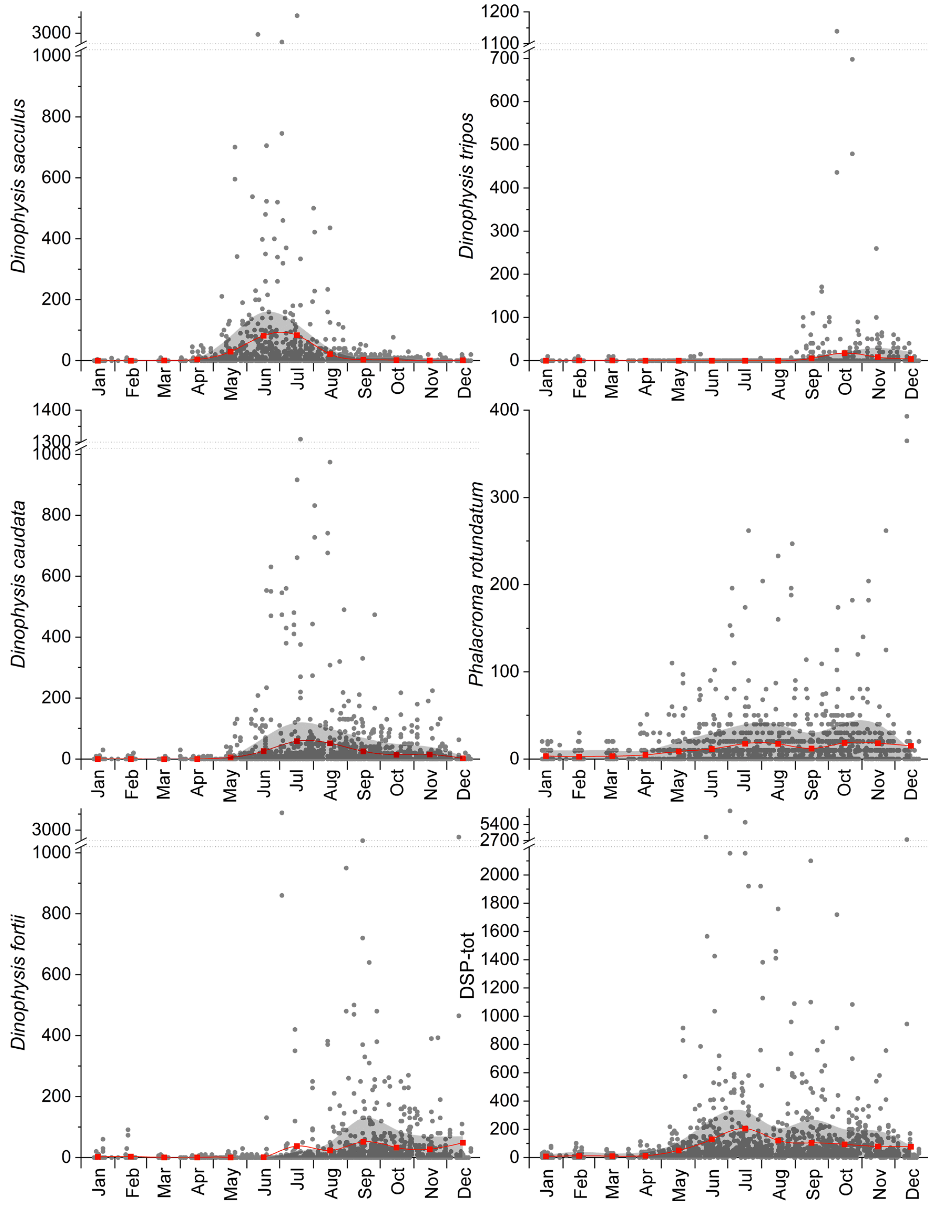


Fig. 4. Annual distribution of abundance (cells l^{-1}) of the five DSP-producing species and DSP-tot used as independent variables in ML models. The dots represent individual observations, the grey area the 90th percentile, while the red line represents the mean for the period 1994–2021. Note the different scaling of the y-axes.

poisoning was lowest from December to April (2% of all positive tests).

3.2. Training ML models and performance evaluation

Using the ML pipeline, SVM, DT, RF and ANN models were trained to predict DSP toxicity, with hyperparameters individually optimised using grid search based on the mean F1 score over 5 folds. Table 2 gives an overview of the best hyperparameters for each ML algorithm.

To obtain a reliable estimate of how the models would perform with similar data in the real world, the entire pipeline (model construction with hyperparameter optimization on the training set and evaluation on the test set) was run through 100 iterations and the results averaged. This was important due to the variability of results between runs and between folds (Fig. 6) and allowed for a more reliable algorithm comparison using the three performance metrics of precision, recall, and F1 score.

Table 3 shows the average performance of the models for each algorithm and performance metric over the 100 iterations. Results

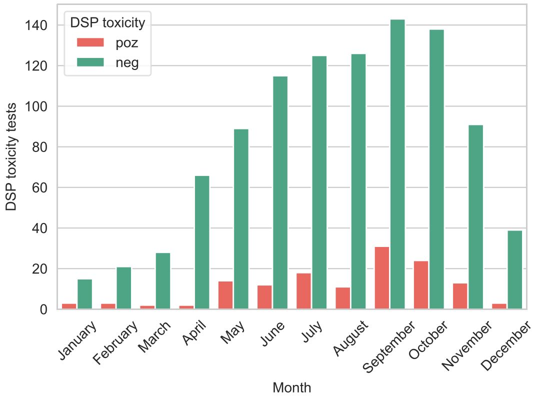


Fig. 5. Monthly number of DSP toxicity tests separated into positive and negative results over the years 1994–2021.

demonstrate that the ML models are able to predict DSP toxicity. RF had the highest and most stable prediction performance over the 100 iterations with an average precision of 0.74 (std ±0.09), recall of 0.59 (std ±0.08) and F1 score of 0.65 (std ±0.07). The study found that among the selected ML algorithms for this particular type of data, RF is the most suitable model for the direct prediction of DSP toxicity in mussels. Interestingly, the less complex DT performed on average very similarly to the ANN, with both having an identical precision of 0.42 and F1 score of 0.43, while the ANN had a slightly higher recall of 0.48 (DT 0.47).

As indicated, the models were optimised for F1 score during parameter tuning, which resulted in slightly lower recall values during evaluation on the test set. Although maximising recall to predict the most positive toxicity cases by optimising for this metric might appear advantageous for real-world EWS, it would also generate many false positives. This is illustrated in the precision-recall curve for the RF model in Fig. 7, where precision sharply declines at recall values above approximately 0.6. Therefore, the study took a balanced approach to recall and precision when evaluating model performance.

3.3. *Explaining machine learning models*

An example of the constructed DT with optimised parameters from grid search (Table 2) and trained on all data (training and test data combined in order to take advantage of all available data) is presented in Fig. 8. In this example, the decisions of the model can be directly interpreted. This gives a good insight into how the model makes predictions. At the first step the model splits the data according to the presence of *D. fortii* at abundances above 30 cells l^{-1}. If the abundance is higher the model predicts positive test results. With abundances equal or lower than 30 cells l^{-1} the decision depends on the presence or absence of *D. caudata*. When *D. caudata* is present, the next decisive variable is salinity, which leads to positive toxicity results at values $\leq$ 36.17 while at higher salinity the prediction is always a negative toxicity result.

To investigate the behaviour of a typical RF model that performed best among the opaque models (F1 score), we applied two explainability methods. Fig. 9 shows the permutation importance, with the variables ordered from top to bottom. The length of each bar indicates how much the model performance decreases when the values of the respective variables are randomly shuffled. In the RF model presented, *D. fortii* is the variable with the greatest influence, followed by DSP-tot and Soča river flow.

Variables in the lower part of the diagram with negative values indicate that the predictions for the shuffled data were more accurate than the original data. However, with small data sets, as is the case here, such a result may occur more frequently due to chance. Therefore, these variables have little or no significance for the predictions of the RF model according to the permutation importance method.

Next, SHAP was used to explain the results of the RF model. In Fig. 10, the beeswarm plot summarises the overall distribution of Shapley values for each variable and example (point) in the test set. A positive Shapley value indicates that the presence of the variable increases the probability of the target class compared to the average prediction. Conversely, a negative Shapley value means that the feature lowers the prediction value. Larger absolute Shapley values indicate that the feature has a stronger effect on the prediction, while a Shapley value close to zero means that the feature has little to no effect. The beeswarm plot provides an overview of which variables are most important for the predictions of the RF model by ordering the variables according to the sum of the absolute Shapley values for each variable in the test set. The distribution of the variable values (colour scale in Fig. 10) along the axis of the Shapley value is important for interpreting the importance of the variables. *D. fortii* is ranked first on the beeswarm plot, which, together with the distribution of red dots on the positive side, indicates that higher abundance of this species influences the prediction of the RF model by increasing confidence in a positive toxicity result. Similar observations can be made for the next two most important variables, *D. caudata* and DSP-tot. In contrast, lower flow rates of the Soča river and higher salinity levels reduce RF confidence in a positive test result. A lower air temperature also appears to influence the model in such a way that the prediction probability for a positive toxicity result decreases. Other variables with Shapley values closer to zero have less influence on the model's decision.

SHAP was also used to explain the model's prediction for a selected example. In Fig. 11, the SHAP force plot illustrates how much each independent variable is contributing to push the output from the base value (0.33) to the actual output produced by the RF model. Variables forcing the prediction higher are shown in red, while those forcing the prediction lower are in blue. As can be seen, RF correctly predicted this example as positive (predicted value = 1) with a confidence of 0.58.

The high abundance of DSP-tot (460 cells l^{-1}), to which *D. fortii* (330 cells l^{-1}) and *D. tripos* (110 cells l^{-1}) contributed the most, and the long duration of solar radiation (187.5 h) have the highest relevance for the prediction result, forcing it to have higher confidence values. In contrast, the low flow rate of the Soča river (1174 m3 s^{-1}) and the high salinity (37.67) are the main variables that reduce the model's confidence in a positive prediction.

Table 2
The best hyperparameters of the ML algorithms determined using grid search.

	SVM	DT	RF	ANN
model	C = 100 Class_weight = None	Class_weight = None criterion=entropy max_depth=4	Class_weight=None criterion=gini N_estimators = 300	activation=relu max_iter=5000 hidden_layer_sizes = 3 batch_size=min(200, n_samples)
SMOTE	k_neighbors=3 sampling_strategy=0.4	k_neighbors=3 sampling_strategy=0.3	k_neighbors=3 sampling_strategy = 0.4	k_neighbors = 5 sampling_strategy = 0.6
RandomUnderSampler	sampling_strategy=0.6	sampling_strategy=0.6	sampling_strategy=0.5	sampling_strategy = 0.7

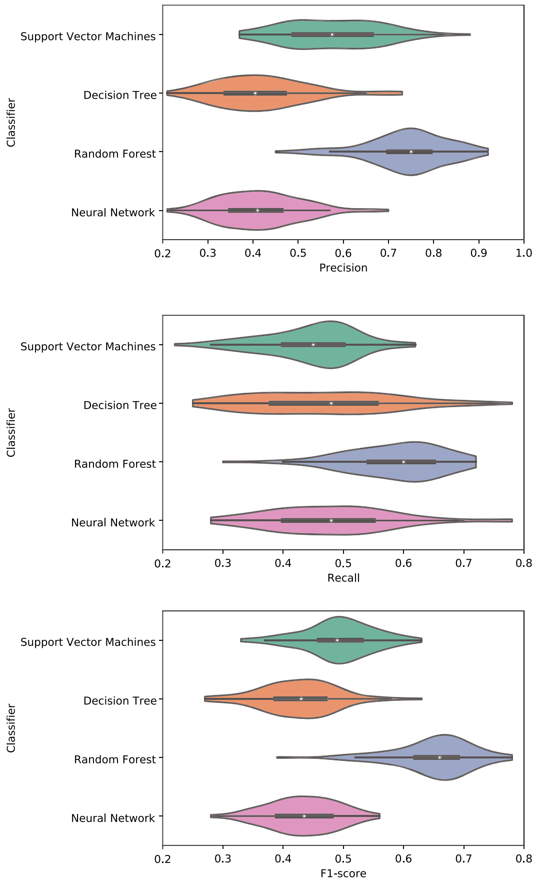


Fig. 6. Variability of performance metrics over 100 iterations.

Table 3
Average model test results for four ML algorithm classes over 100 iterations.

Classifier	Precision	Recall	F1-score
SVM	0.42	0.47	0.43
DT	0.42	0.48	0.43
RF	0.74	0.59	0.65
ANN	0.58	0.45	0.49

4. Discussion

The complexity and non-linearity of the ecological dynamics responsible for DSP events results from a multitude of biological and environmental factors, as well as from the not well understood contamination/decontamination kinetics of mussels (García-Corona et al. 2022). Such an interplay of factors requires advanced predictive models (Cruz et al. 2021) and the right selection of variables (Yu et al. 2021). In this study, several ML algorithms were used for modelling and the five most widely distributed DSP species in Adriatic coastal waters, which pose a significant risk of DSP toxin contamination in mussels, were included (Henigman et al. 2024; Ninčević Gladan et al. 2020). *Dinophysis* species are known to thrive in a stable water column stratification, which is often associated with increased seawater temperature and reduced surface salinity due to precipitation or freshwater discharges. In addition, local hydrodynamic features such as upwelling and downwelling cycles, tides and coastal advection influence *Dinophysis* populations (Reguera et al. 2012). Therefore, variables such as seawater temperature, salinity, turbidity, chlorophyll-a, wind, air temperature and precipitation appeared to be relevant, and they have also been used in other studies to predict toxicity with ML, including DSP (Capoccioni et al. 2023), PSP (Harley et al. 2020), or both (Bouquet et al. 2022). Although not investigated in our study, further physiological aspects of phytoplankton such as toxin production dependent on abiotic factors (e.g. temperature) or intrinsic factors (growth phase) together with bivalve physiology (e.g. bioaccumulation processes) could improve model performance.

The constructed dataset presented several challenges for training ML models. There was a significant class imbalance in the DSP toxicity variable, as events with high toxicity values occurred disproportionately less frequently than events with low values. Consequently, data augmentation techniques were incorporated into the model training pipeline to mitigate this. In addition, dimensionality reduction revealed that instances with positive and negative toxicity values were spread across the dataspace (Fig. 2), suggesting that some of the data sources were noisy (possible reasons include measurement and processing errors as well as changes in monitoring methods during the long time span). Since this causes the ML models to have difficulty in drawing a decision boundary, we removed instances belonging to the negative class whenever they were in close vicinity (in the data space) of instances belonging to the positive class. In this way, the most “conflicting” instances of the majority class were removed and the target classes were further balanced. Although this did not completely eliminate the underlying problem, this approach had a positive effect on the performance of the model.

Despite the challenges, model evaluation demonstrated the effectiveness of ML in predicting DSP toxicity in mussels, with RF algorithms outperforming other models due to their robustness and ability to handle complex interactions within the data (Table 3), which is consistent with findings on the effectiveness of ensemble methods on similar problems (Cruz et al. 2021; Harley et al. 2020). By aggregating the results of multiple decision trees, RF can mitigate overfitting and capture more nuanced patterns. Ensemble methods are also known for their variance reduction and resilience to noise and outliers. In the evaluation, RF showed an average recall of 0.59, with most models exceeding this value (Fig. 6) despite being optimised on the F1 score, which was 0.65. When parameters were optimised, higher recall rates were observed, but for a balanced evaluation, models with better F1 scores were preferred. When implementing EWS, prioritising recall could improve the detection of positive toxicity at the expense of higher FP — a trade-off considered acceptable due to the likely subsequent verification of toxicity in the laboratory.

Both the DT and SVM models exhibited lower F1 scores (0.43). The simple algorithmic nature of DT limits its ability to model the complex relationships while SVM might have struggled with the noisy and imbalanced nature of the data as it performs optimally when there is a clear margin of separation between classes. The ANN model had a slightly higher F1 score (0.49), but still performed worse than RF. ANNs require bigger training sets than other ML algorithms. In our case the ANN was shallow and small leading to limited learning capacity.

The evaluation of the robustness of the models was limited by the small size of the test set, which urges caution when applying these models to different data sources. ML models may have limited generalizability and may perform suboptimal on data that deviate from the training set. In general, it should be kept in mind that site- and species-specific models are superior to generalizable models (Rousso et al. 2020).

By applying XAI approaches in marine ecology, our study aimed to fill important gaps in ML applications in this field while emphasising the need and potential of XAI. Especially for scientific and real-world applications, – such as EWS–, it is insufficient to only estimate how reliable

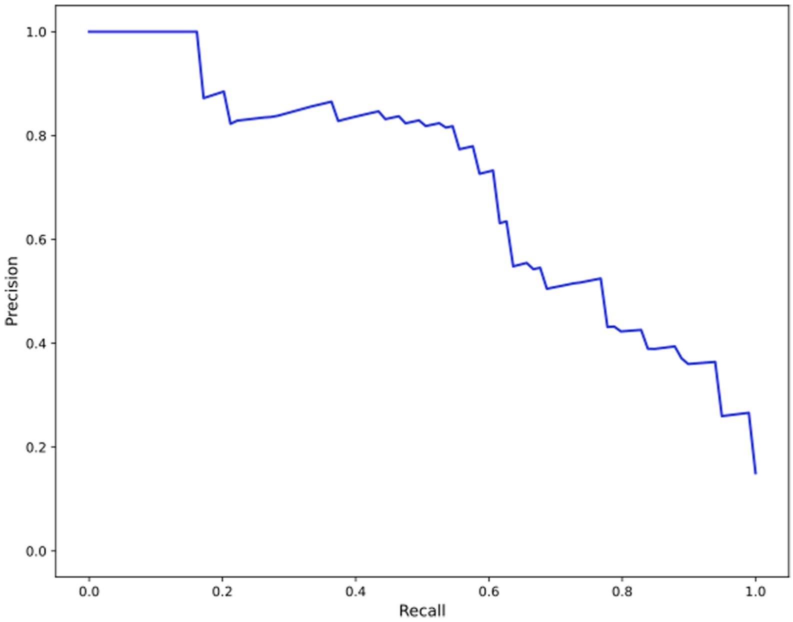


Fig. 7. The precision-recall curve of the best RF model with parameters optimised using grid search.

the predictions of a model are. For an ML-based EWS to be perceived as trustworthy, authorities also need to understand the rationale behind the model's decisions so that they can make an informed decision on appropriate actions.

Our study shows how XAI techniques can provide insights into the model's decisions by revealing the correlations learned by the ML algorithms. In two different explainability methods (Figs. 9 and 10) of the best performing ML model (RF), *D. fortii* was ranked as the most influential variable for positive toxicity prediction, followed by the entire DSP assemblage or other species such as *D. caudata*. These results are largely consistent with real data from in vitro studies demonstrating the synthesis of DSP toxins in isolated cells of *D. fortii* (Yasumoto et al. 1980; M. G. Park et al. 2006) or from field studies. Our data show that the period with the highest abundance of *D. fortii* coincides with the periods with the highest incidence of DSP above the regulatory limit in September and October (Figs. 4 and 5), while earlier DSP events, from May to August, could be due to *D. caudata* or *D. sacculus*, which also increased the probability of positive test results. In addition, a recent study from Slovenian coastal waters, conducted with a much shorter dataset (6 years), showed the strongest association between DSP toxins in mussels and *D. fortii* (Henigman et al. 2024).

The next highly ranked variables are those related to the supply of freshwater to the coastal waters, such as the flow rate of the Soča River and precipitation. These variables alter the hydrodynamics and increase the stratification of the water column by reducing salinity in addition to the high water temperatures during warm periods (Boicourt et al. 2020). Such conditions can promote the growth of *Dinophysis* species and their accumulation in thin layers (Broullón et al. 2020; Harred and Campbell 2014). The interplay between these influential abiotic factors and the dynamics of DSP species, particularly *D. fortii*, is also evidenced by a real case of mussel toxicity that was correctly predicted by the model with relatively high confidence (Fig. 11). The factors that increased the predictive confidence of the model the most were the high abundance of *D. fortii* and DSP-tot (330 cells l^{-1} and 460 cells l^{-1} respectively are among the highest values in our study area) and the long day length, serving as a proxy for warm, stable summer conditions. Conversely, high salinity, low cumulative river flow and the month of September had the opposite effect and reduced confidence in the model's positive prediction. While the first two factors were correctly identified as least influential for a positive prediction, the influence of the month led to a bias in the model result, as the highest number of toxic events occurred in September (see Fig. 5). Although the drivers mentioned above clearly influenced the predictions of the model, this does not prove causality as correlation does not imply causation.

The abiotic variables highlighted as important in the present study are in good agreement with the results of previous studies in this area, where water temperatures and salinity were most responsible for the temporal variability and high abundances of DSP-producing species (France and Mozetic 2006; Henigman et al. 2024). Besides, the results of the ML models in the present study revealed some other important variables, such as river flow and precipitation, suggesting not only their influence on pycnocline formation, but also possibly the relationship with nutrients.

Although the DT model exhibits lower performance than the RF model, it also identifies a similar set of variables as RF. However, DT offers better interpretability for potential end users due to its simplicity and ease of use. Our example shows (Fig. 8) that even low abundances of *D. fortii* (30 cells l^{-1} or more) can serve as warning indicators of positive toxicity. In cases where *D. fortii* has <30 cells l^{-1}, the presence of *D. caudata* and salinity levels ≤ 36.17 should be considered as warning indicators. However, these thresholds have to be taken with caution since the models were trained with abundance data of DSP-producing species obtained with different sampling methodologies. Despite its suboptimal performance, DT provides thresholds for decisions that can be easily verified retrospectively. When applied to the data of a mussel farm, the informative value was significantly improved from 2013, when the more precise analytical method for the detection of DSP toxins was introduced into the Slovenian national monitoring of mussels.

5. Conclusions

The study represents a step forward in the less researched and more

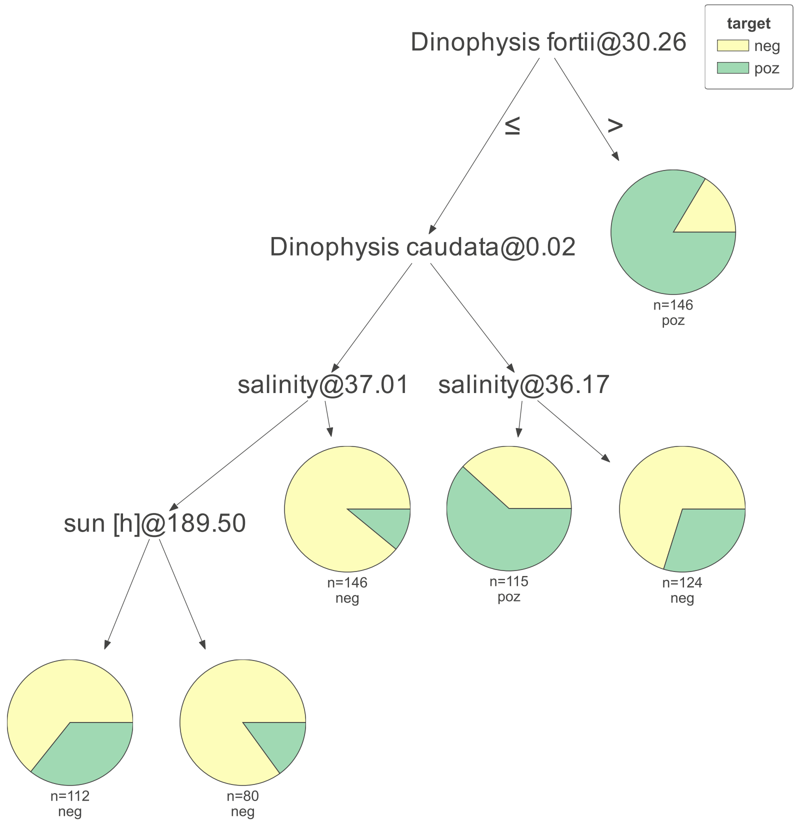


Fig. 8. Visualisation of a constructed DT model to illustrate the decision rules for DSP toxicity results. The numbers at each leaf node (pie chart) indicate the total number of remaining instances. The prediction of the model is labelled "poz" for positive toxicity predictions and "neg" for negative toxicity predictions.

difficult problem of using ML to directly predict DSP toxicity in mussels in the Adriatic Sea. To this end, a comprehensive dataset of toxic phytoplankton and DSP toxins from the GoT, spanning three decades has been created and is openly available for further research. The focus of the study is the application of XAI principles by using interpretable ML models and explainability methods for opaque models to gain helpful insights into the complex interactions between marine organisms and their environment.

The following main conclusions can be drawn from the study:

- The right data preprocessing steps are crucial for overcoming the specific challenges of consolidated datasets from different sources and determine the training and performance of the model.
- ML models, especially RF, can satisfactorily predict DSP toxicity in mussels from Slovenian mussel farms.
- Both RF explainability methods and DT visualisations show that *Dinophysis fortii* and *Dinophysis caudata* together with abiotic factors influencing salinity of coastal waters (river discharge and precipitation) have the greatest influence on the predictions.
- The knowledge gained through explainability methods about the behaviour of the model makes ML approaches suitable for the EWS due to the increased trustworthiness.
- Predictive performance should be calibrated for the needs of EWS by optimising model training on the performance metric that is most important in the real world (e.g., recall).

The ML models developed, when integrated into EWS, can provide a cost-effective means of implementing timely and appropriate mitigation actions, such as trade bans, while improving management strategies to minimise health risks and social and economic damage. Future improvements should focus on refining these models by expanding the training dataset, especially with more positive toxicity tests, and by continuously improving the temporal resolution and quality of the training data through improved monitoring methods. For example, the robustness of these models could be significantly improved by including a wider range of data of high-toxicity events and incorporating datasets from neighbouring regions.

CRediT authorship contribution statement

Martin Marzidovšek: Writing – review & editing, Writing – original draft, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Janja Francé:** Writing – review & editing, Writing – original draft, Investigation, Data curation, Conceptualization. **Vid Podpečan:** Writing – review & editing, Writing – original draft, Supervision, Software, Methodology, Formal analysis, Conceptualization. **Stanka Vadnjal:** Writing – review & editing, Investigation. **Jožica Dolenc:** Writing – review & editing, Investigation. **Patricija Mozetič:** Writing – review & editing, Writing – original draft, Investigation, Conceptualization.

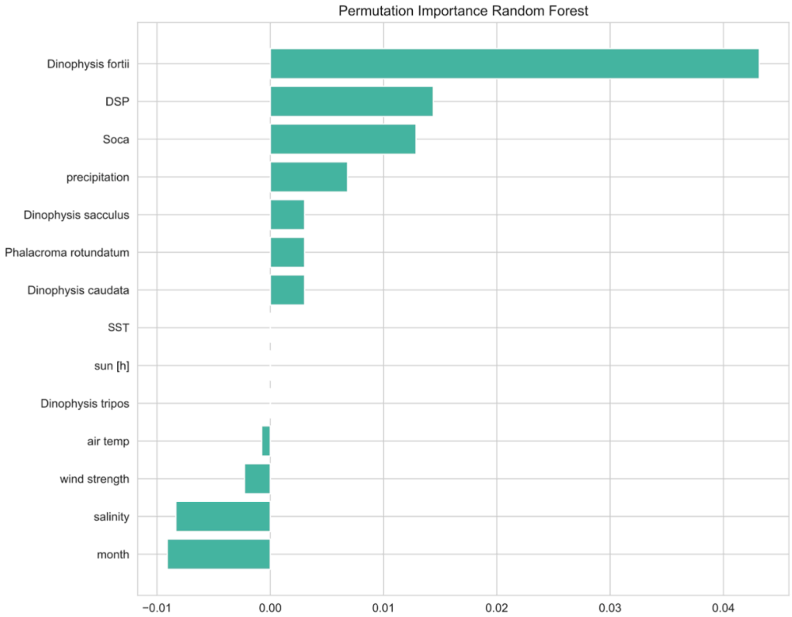


Fig. 9. Permutation importance of a typical RF model showing the ranking of the relevance of the independent variables for the prediction of DSP toxicity.

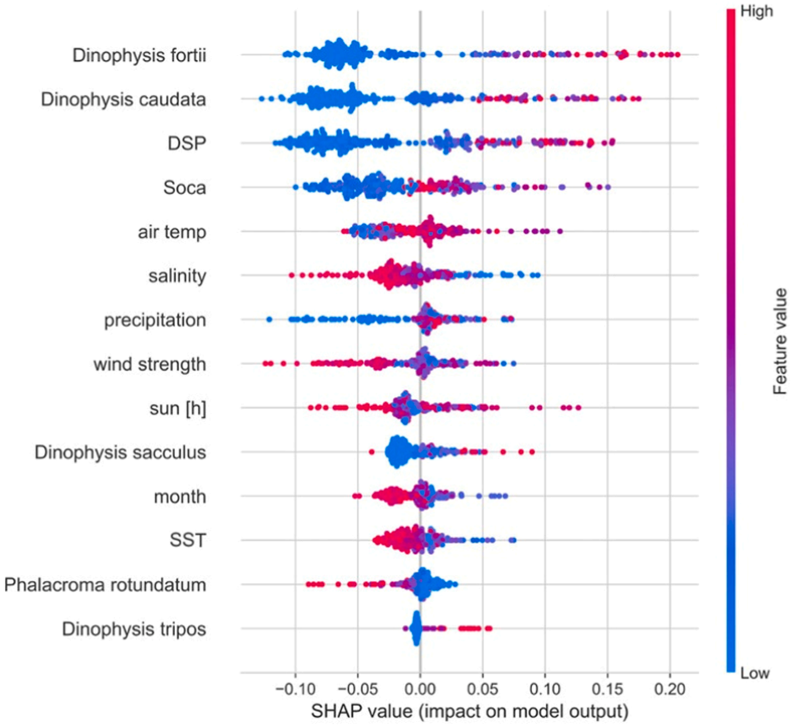


Fig. 10. The summary beeswarm plot of the Shapley values for the independent variables of each instance in the test set predicted by a typical RF model.

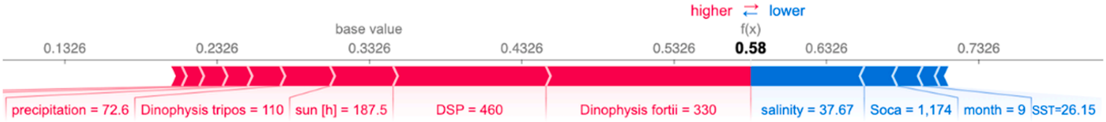


Fig. 11. SHAP force plot of individual RF model prediction of DSP toxicity (names of some less impactful variables on the left – air temp, D. caudata and Phalacroma rotundatum – are not shown).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Online repository

Acknowledgments

This work was financially supported by the National monitoring program for toxic phytoplankton and marine biotoxins in shellfish growing areas in the Slovenian Sea (The Administration of the Republic of Slovenia for Food Safety, Veterinary Sector and Plant Protection) and by the Slovenian Research and Innovation Agency (grant numbers P1–0237, P4–0092 and P2–0103).

References

Accoroni, Stefano, Cangini, Monica, Angeletti, Roberto, Losasso, Carmen, Bacchiocchi, Simone, Costa, Antonella, Di Taranto, Aurelia, et al., 2024. Marine Phycotoxin levels in shellfish—14 years of data gathered along the italian coast. Harmful Algae 131 (January), 102560.

Avdelas, Lamprakis, Avdic-Mravlje, Edo, Borges Marques, Ana Cristina, Cano, Suzana, Capelle, Jacob J., Carvalho, Natacha, Cozzolino, Maria, et al., 2021. The decline of mussel aquaculture in the European Union: causes, economic impacts and opportunities. Rev. Aquac. 13 (1), 91–118.

Boicourt, William C., Ličer, Matjaž, Li, Ming, Vodopivec, Martin, Malačič, Vlado, 2020. Sea State. Coastal Ecosystems in Transition. Wiley. https://doi.org/10.1002/9781119543626.ch3.

Bouquet, Aurélien, Laabir, Mohamed, Rolland, Jean Luc, Chomérat, Nicolas, Reynes, Christelle, Sabatier, Robert, Felix, Christine, Berteau, Tom, Chiantella, Claude, Abadie, Eric, 2022. Prediction of Alexandrium and Dinophysis algal blooms and shellfish contamination in french mediterranean lagoons using decision trees and linear regression: a result of 10 years of sanitary monitoring. Harmful Algae 115 (102234), 102234.

Breiman, Leo., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Broullón, Esperanza, López-Mozos, Marta, Reguera, Beatriz, Chouciño, Paloma, Doval, María Dolores, Fernández-Castro, Bieito, Gilcoto, Miguel, Nogueira, Enrique, Souto, Carlos, Mouriño-Carballido, Beatriz, 2020. Thin layers of phytoplankton and harmful algae events in a coastal upwelling system. Prog. Oceanogr. 189 (November), 102449.

Brush, Mark J., Mozetič, Patricija, Francé, Janja, Bernardi Aubry, Fabrizio, Djakovac, Tamara, Faganeli, Jadran, Harris, Lora A., Niesen, Meghann, 2020. Phytoplankton dynamics in a changing environment. Coastal Ecosystems in Transition. Wiley. https://doi.org/10.1002/9781119543626.ch4.

Cantoni, Carolina, Cozzi, Stefano, Pecchiar, Irene, Cabrini, Marina, Mozetič, Patricija, Catalano, Giulio, Umani, Serena Fonda, 2003. Short-term variability of primary production and inorganic nitrogen uptake related to the environmental conditions in a shallow coastal Area (Gulf of Trieste, N Adriatic Sea). Oceanologica Acta 26 (5), 565–575.

Capoccioni, Fabrizio, Bille, Laura, Colombo, Federica, Contiero, Lidia, Martini, Arianna, Mattia, Carmine, Napolitano, Riccardo, Tonachella, Nicolò, Toson, Marica, Pulcini, Domitilla, 2023. A predictive model for the bioaccumulation of Okadaic acid in Mytilus Galloprovincialis farmed in the Northern Adriatic Sea: a tool to reduce product losses and improve mussel farming sustainability. Sustainability: Sci. Practice Policy 15 (11), 8608.

Ciminiello, Patrizia, Dell'Aversano, Carmela, Fattorusso, Ernesto, Forino, Martino, Magno, Giuseppina S., Tartaglione, Luciana, Quilliam, Michael A., Tubaro, Aurelia, Poletti, Roberto, 2005. Hydrophilic interaction liquid chromatography/mass spectrometry for determination of Domoic acid in Adriatic shellfish. Rapid Commun. Mass Spectrom.: RCM 19 (14), 2030–2038.

Comici, Cinzia and Bussani, Andrea. 2007. “Analysis of the River Isonzo Discharge (1998-2005).” https://bgo.ogs.it/sites/default/files/pdf/bgta48.4_COMICI.pdf.

Cozzi, Stefano, Falconi, Claus, Comici, Cinzia, Čermelj, Branko, Kovac, Nives, Turk, Valentina, Giani, Michele, 2012. Recent evolution of river discharges in the gulf of trieste and their potential response to climate changes and anthropogenic pressure. Estuar. Coast. Shelf. Sci. 115 (December), 14–24.

Cozzi, Stefano, Giani, Michele, 2011. River water and nutrient discharges in the northern adriatic sea: current importance and long term changes. Cont. Shelf. Res. 31 (18), 1881–1893.

Council of the European Communities, 1991. Council Directive of 15 July 1991 laying down the health conditions for the production and the placing on the market of live bivalve molluscs (91/492/EEC). Off. J. Eur. Commun. 268, 1–14.

Cruz, Rafaela C., Costa, Pedro Reis, Vinga, Susana, Krippahl, Ludwig, Lopes, Marta B., 2021. A review of recent machine learning advances for forecasting harmful algal blooms and shellfish contamination. J. Mar. Sci. Eng. 9 (3), 283.

Derot, Jonathan, Yajima, Hiroshi, Jacquet, Stéphan, 2020. Advances in forecasting harmful algal blooms using machine learning models: a case study with Planktothrix Rubescens in Lake Geneva. Harmful Algae 99 (November), 101906.

Falcieri, Francesco Marcello, Kantha, Lakshmi, Benetazzo, Alvise, Bergamasco, Andrea, Bonaldo, Davide, Barbariol, Francesco, Malačič, Vlado, Sclavo, Mauro, Carniel, Sandro, 2016. Turbulence observations in the gulf of trieste under moderate wind forcing and different water column stratification. Ocean Sci. 12 (2), 433–449.

France, Janja, Mozetic, Patricija, 2006. Ecological characterization of toxic phytoplankton Species (Dinophysis Spp., Dinophyceae) in slovenian Mariculture areas (Gulf of Trieste, Adriatic Sea) and the Implications for monitoring. Mar. Pollut. Bull. 52 (11), 1504–1516.

García-Corona, José Luis, Hégaret, Hélène, Deléglise, Margot, Marzari, Adeline, Rodríguez-Jaramillo, Carmen, Foulon, Valentin, Fabioux, Caroline, 2022. First subcellular localization of the amnesic shellfish toxin, Domoic acid, in bivalve tissues: deciphering the physiological mechanisms involved in its long-retention in the king scallop Pecten maximus. Harmful Algae 116 (July), 102251.

Grasso, Isabella, Archer, Stephen D., Burnell, Craig, Tupper, Benjamin, Rauschenberg, Carlton, Kanwit, Kohl, Record, Nicholas R., 2019. The hunt for red tides: deep learning algorithm forecasts shellfish toxicity at site scales in coastal maine. Ecosphere 10 (12). https://doi.org/10.1002/ecs2.2960.

Guallar, Carles, Delgado, Maximino, Diogène, Jorge, Fernández-Tejedor, Margarita, 2016. Artificial neural network approach to population dynamics of harmful algal blooms in Alfacs bay (NW Mediterranean): case studies of Karlodinium and pseudo-Nitzschia. Ecol. Modell. 338 (October), 37–50.

Hallegraeff, Gustaaf M., Anderson, Donald M., Belin, Catherine, Bottein, Marie-Yasmine, Bresnan, Eileen, Chinain, Mireille, Enevoldsen, Henrik, et al., 2021. Perceived global increase in algal blooms is attributable to intensified monitoring and emerging bloom impacts. Commun. Earth. Environ. 2 (June). https://doi.org/10.1038/s43247-021-00178-8.

Harley, John R., Lanphier, Kari, Kennedy, Esther, Whitehead, Chris, Bidlack, Allison, 2020. Random forest classification to determine environmental drivers and forecast paralytic shellfish toxins in southeast alaska with high temporal resolution. Harmful Algae 99 (November), 101918.

Harred, Laura Brooke, Campbell, Lisa, 2014. Predicting harmful algal blooms: a case study with Dinophysis ovum in the gulf of Mexico. J. Plankton Res. 36 (6), 1434–1445.

Henigman, Urška, Mozetič, Patricija, Francé, Janja, Knific, Tanja, Vadnjal, Stanka, Dolenc, Jožica, Kirbiš, Andrej, Biasizzo, Majda, 2024. Okadaic acid as a major problem for the seafood safety (Mytilus Galloprovincialis) and the dynamics of toxic phytoplankton in the Slovenian coastal sea (Gulf of Trieste, Adriatic Sea). Harmful Algae 135 (May), 102632.

Kim, Jin Hwi, Jae-Ki, Shin, Hankyu, Lee, Dong, Hoon Lee, Joo-Hyon, Kang, Kyung, Hwa Cho, Yong-Gu, Lee, Kangmin, Chon, Sang-Soo, Baek, Yongeun, Park, 2021. Improving the performance of machine learning models for early warning of harmful algal blooms using an adaptive synthetic sampling method. Water Res. 207 (December), 117821.

Lemaître, Guillaume, Nogueira, Fernando, Aridas, Christos K., 2017. Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. J. Mach. Learn. Res.: JMLR 18 (17), 1–5.

Liu, Yao, Qiao, Fu, Xu, Lele, Wang, Runtao, Jiang, Wei, Xu, Zhen, 2022. Fast detection of Diarrhetic shellfish poisoning toxins in mussels using NIR spectroscopy and improved twin support vector machines. Front. Mar. Sci. 9. https://doi.org/10.3389/fmars.2022.907378.

Lundberg, Scott M., Lee, Su-In, 2017. A Unified Approach to Interpreting Model Predictions. Adv Neural Inf Process Syst 30. In: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

Malačič, Vlado., 1991. Estimation of the vertical eddy diffusion-coefficient of heat in the gulf of trieste (northern Adriatic). Oceanologica Acta 14, 23–32.

McInnes, Leland, Healy, John, Melville, James, 2018. UMAP: uniform manifold approximation and projection. J. Open. Source Softw. 3 (29) (861).

McKinney, Wes, and Others, 2010. Data structures for statistical computing in Python. In: Proceedings of the 9th Python in Science Conference, 445, pp. 51–56. Austin, TX.

Menardi, Giovanna, Torelli, Nicola, 2014. Training and assessing classification rules with imbalanced data. Data Min. Knowl. Discov. 28 (1), 92–122.

Mosetti, Ferruccio, Purga, Nicolo., 1990. Courants côtiers de différente origine dans un petit golfe (Golfe de Trieste). Boll. Oceanol. Teor. Appl 8, 51–62.

Mozetič, Patricija, Francé, Janja, Kogovšek, Tjaša, Talaber, Iva, Malej, Alenka, 2012. Plankton trends and community changes in a coastal Sea (northern Adriatic): bottom-up vs. top-down control in relation to environmental drivers. Estuar. Coast. Shelf Sci. 115 (December), 138–148.

Muttil, Nitin, Chau, Kwok-Wing, 2006. Neural network and genetic programming for modelling coastal algal blooms. Int. J. Environ. Pollut. 28 (3–4), 223–238.

Gladan, Ninčević, Frano Matić, Živana, Arapov, Jasna, Skejić, Sanda, Bužančić, Mia, Bakrač, Ana, Straka, Maja, et al., 2020. The relationship between toxic phytoplankton species occurrence and environmental and meteorological factors along the eastern Adriatic coast. Harmful Algae 92 (February), 101745.

Gladan, Nincevic, Ivana Ujevic, Zivana, Milandri, Anna, Marasovic, Ivona, Ceredi, Alfiero, Pigozzi, Silvia, Arapov, Jasna, Skejic, Sanda, 2011. Lipophilic toxin profile in Mytilus Galloprovincialis during Episodes of Diarrhetic Shellfish poisoning (DSP) in the N.E. Adriatic Sea in 2006. Molecules. 16 (1), 888–899.

Park, Myung Gil, Kim, Sunju, Kim, Hyung Seop, Myung, Geumog, Kang, Yi Gu, Yih, Wonho, 2006. First successful culture of the marine dinoflagellate Dinophysis acuminata. Aquat. Microb. Ecol. 45 (2), 101–106.

Park, Yongeun, Lee, Han Kyu, Shin, Jae-Ki, Chon, Kangmin, Kim, Sunghwan, Cho, Kyung Hwa, Kim, Jin Hwi, Baek, Sang-Soo, 2021. A machine learning approach for early warning of Cyanobacterial bloom outbreaks in a freshwater reservoir. J. Environ. Manage. 288 (June), 112415.

Patrício, André, Lopes, Marta B., Costa, Pedro Reis, Costa, Rafael S., Henriques, Rui, Vinga, Susana, 2022. Time-lagged correlation analysis of shellfish toxicity reveals predictive links to adjacent areas, species, and environmental conditions. Toxins. (Basel)Toxins. (Basel) 14 (10). https://doi.org/10.3390/toxins14100679.

Pedregosa, Fabian, Varoquaux, Gael, Gramfort, Alexandre, Michel, Vincent, Thirion, Bertrand, Grisel, Olivier, Blondel, Mathieu, et al., 2011. Scikit-Learn: machine learn-Ing in Python. J. Mach. Learn. Res. 12.

Prassopoulou, Eleanna, Katikou, Panagiota, Georgantelis, Dimitrios, Kyritsakis, Apostolos, 2009. Detection of Okadaic acid and related esters in mussels during Diarrhetic shellfish poisoning (DSP) episodes in greece using the mouse bioassay, the PP2A inhibition assay and HPLC with Fluorimetric detection. Toxicon: Off. J. Int. Soc. Toxinol. 53 (2), 214–227.

Recknagel, Friedrich., 1997. ANNA–artificial neural network model for predicting species abundance and succession of blue-green algae. Hydrobiologia 349 (1/3), 47–57.

Recknagel, Friedrich, Orr, Philip T., Cao, Hongqing, 2014. Inductive reasoning and forecasting of population dynamics of Cylindrospermopsis Raciborskii in three sub-tropical reservoirs by evolutionary computation. Harmful Algae 31 (January), 26–34.

Reguera, Beatriz, Velo-Suárez, Lourdes, Raine, Robin, Park, Myung Gil, 2012. Harmful Dinophysis species: a review. Harmful Algae 14 (February), 87–106.

Ribeiro, Marco Tulio, Singh, Sameer, Guestrin, Carlos, 2016. 'Why should I Trust You?': explaining the predictions of any classifier. *arXiv [cs.LG]*. arXiv. https://doi.org/10.1145/2939672.2939778.

Rousso, Benny Zuse, Bertone, Edoardo, Stewart, Rodney, Hamilton, David P., 2020. A systematic literature review of forecasting and predictive models for cyanobacteria blooms in freshwater lakes. Water Res. https://doi.org/10.1016/j.watres.2020.115959.

Sagarminaga, Yolanda, Garcés, Esther, Francé, Janja, Stern, Rowena, Revilla, Marta, Magaletti, Erika, Bresnan, Eileen, et al., 2023. New tools and recommendations for a better management of harmful algal blooms under the european marine strategy framework directive. Front. Ocean Sustain. 1. https://doi.org/10.3389/focsu.2023.1298800.

Shimoda, Yuko, Arhonditsis, George B., 2016. Phytoplankton functional type modelling: running before we can walk? A critical evaluation of the current state of knowledge. Ecol. Modell. 320 (January), 29–43.

Stravisi, Franco. "Some characteristics of the circulation in the Gulf of Trieste." (1983). https://ricerca.unityfvg.it/entities/publication/3d43cf49-583a-40a6-a35b-a53282302725.

Ujević, Ivana, Roje, Romana, Živana, Ninčević-Gladan, Ivona, Marasović, 2012. First report of paralytic shellfish poisoning (PSP) in mussels (Mytilus Galloprovincialis) from Eastern Adriatic Sea (Croatia). Food Control 25 (1), 285–291.

Utermöhl, Hans., 1958. Zur Vervollkommnung der Quantitativen Phytoplankton-Methodik. Schweizerbart.

Yasumoto, Takeshi, Oshima, Yasukatsu, Sugawara, Wataru, Fukuyo, Yasuwo, Oguri, Hajime, Igarashi, Teruo, Fujita, Noritaka, 1980. Identification of *Dinophysis Fortii* as the causative organism of Diarrhetic shellfish poisoning. Nippon Suisan. Gakkaishi 46 (11), 1405–1411.

Yasumoto, Takeshi, Oshima, Yasukatsu, Yamaguchi, Magoichi, 1978. Occurrence of a new type of shellfish poisoning in the Tohoku district. Nippon Suisan Gakkaishi 44 (11), 1249–1255.

Yu, Qiuyan, Ji, Wenjie, Prihodko, Lara, Ross, C.Wade, Anchang, Julius Y., Hanan, Niall P., 2021. Study becomes insight: ecological learning from machine learning. Methods Ecol. Evolut./Brit. Ecol. Soc. 12 (11), 2117–2128.

Zingone, Adriana, Escalera, Laura, Aligizaki, Katerina, Fernández-Tejedor, Margarita, Ismael, Amany, Montresor, Marina, Mozetič, Patricija, Taş, Seyfettin, Totti, Cecilia, 2021. Toxic marine microalgae and noxious blooms in the mediterranean sea: a contribution to the global HAB status report. Harmful Algae 102 (February), 101843.

Chapter 3

Unsupervised Machine Learning in Agricultural Science

This chapter presents the second set of contributions to the dissertation related to the application of explainable ML in agriculture. The aim was to identify the barriers and incentives to the use of DSS for IPM, which is critical for mitigating the negative impacts of agriculture on natural ecosystems and biodiversity. The effectiveness of IPM could be significantly improved by increasing the use of DSS, which are not yet widely used by farmers and farm advisors in Europe. To identify barriers and incentives for the adoption and use of DSS in IPM in Europe, 100 farmers and 119 farm advisors from 12 European countries were surveyed in 2020.

My contribution to this research was the use of unsupervised ML on the tabular data. Therefore, only the methodology and results related to this part are presented here, while the full work is presented in the publication below. In the first step, a dimensionality reduction method was applied to the questionnaire response data using the Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018). Dimensionality reduction was performed separately for both datasets (i.e. numeric/ordinal and nominal attributes) using appropriate distance metrics: Euclidean for numeric/ordinal and Hamming for binary attributes (nominally coded). Before proceeding with clustering, the obtained low-dimensional embeddings of the two datasets were combined by multiplication (intersection).

In the second step, a density-based clustering method was used to objectively test for the presence of clusters, as the integrated response profiles of the participants in the previous step showed a tendency towards clustering. The HDBSCAN algorithm (Campello et al., 2013) was used to cluster the survey participants in the x, y plane and automatically determine the most appropriate number of clusters. Due to the unsupervised learning approach, there was no ground truth to measure the classification error, but the obtained clusters were evaluated using the density-based clustering validation (DBCV) (Moulavi et al., 2014), which gives an indication of how well the clusters are separated from each other. The results were visualised for better understanding and greater trust in the found clusters.

Dimensionality reduction and clustering were performed in Python within JupyterLab, the latest interface of the Jupyter project (Loizides & Schmidt, 2016), using the scikit-learn ML library (Pedregosa et al., 2012), pandas (McKinney et al., 2010), density-based clustering validation (DBCV) (Moulavi et al., 2014), HDBSCAN algorithm (Campello et al., 2013), and for visualisation seaborn (Waskom, 2021) and Matplotlib (Hunter, 2007).

The results show the following. The cluster analysis provided insights into the variability of survey respondents' answers regarding their experiences and expectations of using

DSS in IPM. Both farmers and farm advisors show a tendency to form three groups. Outliers were found among both farmers and farm advisors that could not be assigned to any of the detected clusters by the HDBSCAN algorithm. The discovered clusters in the data of farmers from all participating countries reached a DBCV score of 0.25, while the data of farm advisors in the same countries reached a DBCV score of -0.12. Further analyses based on my work then revealed what were the most common barriers and incentives to access and uptake of DSS by farmers and farm advisors characteristic of each of the identified clusters. These findings will allow experts to direct future research towards finding appropriate solutions to overcome these barriers.

This work directly addresses the following hypothesis defined in the introduction:

H3: Unsupervised ML and explainability methods can reveal user attitudes towards utilising DSS in IPM.

Hypothesis H3 is fully confirmed by the results presented in this chapter.

The research work is presented in the following publication and is included in the following:

Marinko, J., Ivanovska, A., Marzidovšek, M., Ramsden, M., & Debeljak, M. (2023). Incentives and barriers to adoption of decision support systems in integrated pest management among farmers and farm advisors in Europe. *International Journal of Pest Management*, 1-18.

Martin Marzidovšek contributed to this work as follows: conceptualisation, methodology, software, formal analysis, writing the original draft, reviewing and editing, visualisation.

INTERNATIONAL JOURNAL OF PEST MANAGEMENT
https://doi.org/10.1080/09670874.2023.2244912

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

Incentives and barriers to adoption of decision support systems in integrated pest management among farmers and farm advisors in Europe

Jurij Marinko[a,b], Aneta Ivanovska[b,c], Martin Marzidovšek[b], Mark Ramsden[d] and Marko Debeljak[a,b]

[a]Jozef Stefan International Postgraduate School, Ljubljana, Slovenia; [b]Department of Knowledge Technologies, Jozef Stefan Institute, Ljubljana, Slovenia; [c]Faculty of Management and Engineering, University of Nova Gorica, Vipava, Slovenia; [d]RSK ADAS Ltd, Cambridge, UK

ABSTRACT
The effectiveness of integrated pest management (IPM) could be improved by using decision support systems (DSS). To identify barriers and incentives to the adoption of DSS in IPM in Europe, we conducted a survey among farmers and farm advisors in 12 European countries. The data were analysed using state-of-the-art machine learning methods complemented by statistical analysis. The results reveal region-specific barriers to DSS adoption among both farmers and advisors. Among farmers, lack of trust in DSS and the perceived need for additional information technology training are the predominant barriers, while among advisors, limited access to information about DSS is predominant.

ARTICLE HISTORY
Received 22 April 2022
Accepted 1 August 2023

KEYWORDS
Integrated pest management; European scale; decision support system; machine learning; barriers; user adoption

1. Introduction

Integrated Pest Management (IPM) is a science-based practice for the identification and management of pests through a combination of approaches, including biological, cultural, mechanical and chemical measures (Dara 2019). The use of non-chemical measures should be prioritised when possible (Barzman et al. 2015; Stenberg 2017). However, chemical treatment may be used preventively if needed (Dara 2019). Since 2014, farmers in the European Union (EU) have been obliged to follow the guidelines of IPM under European legislation (Directive 2009/128/EC), through which the EU aims to achieve judicious and sustainable use of pesticides.

With rising demand for high-quality, sustainably produced food (European Commission 2020), farmers and farm advisors face increasingly complex decision-making problems on a daily basis (Hernandez et al. 2017). The decision-making process is also becoming increasingly difficult due to various factors such as climate change (Deutsch et al. 2018; Eitzinger, Binder, and Meyer 2018; Skendžić et al. 2021), evolving regulatory and market policies, and technological advancements (e.g. information technologies (IT)) (Lesecq et al. 2022). To help manage complex decisions, numerous decision support systems (DSS) have been developed in Europe for IPM (Mahaman et al. 2003; Tamayo, Ibarra, and Macías 2010; Steffek et al. 2011; Sudharsan et al. 2012; Damos 2015; Marić, Orović, and Stanković 2016; Sciarretta et al. 2019). Advanced IPM DSS that incorporate existing and validated knowledge from different fields such as, agronomy, meteorology, pedology, entomology, phytomedicine and ecology (Taechatanasat and Armstrong 2014), as well as information technology and computer science provide accurate and efficient support for pest and disease management (Demirel and Kumral 2021; Lázaro, Makowski, and Vicent 2021; Maraveas 2022).

Unfortunately, there are several unresolved challenges to the adoption and implementation of DSS in IPM, including the need to increase trust in DSS; provide validated, relevant results to users with different needs, ensure access to information about different developed DSS, and ensure a successful user experience and application of DSS. Due to these and many other challenges, there is a large gap between the number of DSS that have been developed, and the number in use on farms (Gent et al. 2013; Rose et al. 2018). Addressing this gap would lead to more effective implementation of sustainable practices in the use of pesticides (Lázaro, Makowski, and Vicent

CONTACT Jurij Marinko jurij.marinko@ijs.si

Supplemental data for this article can be accessed online at https://doi.org/10.1080/09670874.2023.2244912.

2021) and thus have a positive impact on the health of pesticide users (reduced exposure), the environment (reduced pollution of soil, surface water and groundwater by pesticides) and the ecosystem (reduced impact of pesticides on non-target organisms), (Ali et al. 2021).

Some DSS use mathematical decision theory to help users evaluate their decision alternatives in a consistent manner (Introne and Iandoli 2014) and find the most appropriate step or set of steps to optimally solve a given decision problem (Ara et al. 2021). Their use in agriculture can help solve many problems, such as verifying the effectiveness of implementing sustainable farm practises for pest and disease control (Newman, Lynch, and Plummer 2000). The use of such systems should therefore be more widely adopted by farmers and farm advisors in Europe and globally.

Previous studies have highlighted a lack of trust between developers and DSS users as a cause for the low uptake and use of DSS in agriculture at a local or national scale, along with an inherent risk aversion within user communities (Hochman and Carberry 2011; Shtienberg 2013). These are embedded in a fragmentation of developer and user communities, limiting the ability of users to assess the benefits and scientific basis of DSS and the ability of developers to adequately address the perceived/actual risks experienced by target end users (Gent, De Wolf, and Pethybridge 2011; Rose et al. 2016). In this work, we have extended the research to the European scale and regulatory content, recognising the cross-border nature of many key pest species, and the need for intergovernmental consistency on methods for promoting and supporting IPM. At the same time, there are some region-specific environmental, economic and social specifics within Europe that must be considered. To better understand the users of DSS in agriculture and in IPM at the European level, we conducted a survey among farmers and farm advisors about their needs, experiences and expectations regarding the use of DSS in IPM.

The work presented in this paper was part of the EU project IPM Decisions (grant number 817617, https://www.ipmdecisions.net) with an objective to link IPM DSS from across Europe into a user-friendly platform (http://www.platform.ipmdecisions.net/) that allows easier access and possibly wider adoption of such systems. To increase adoption of IPM DSS, the identification of barriers and incentives is crucial. The aim of our study was to identify and explain (i) the attributes describing the reasons why farmers and farm advisors in Europe use IPM DSS, (ii) the reasons why farmers and farm advisors do not use IPM DSS, and (iii) whether there are differences between the above reasons in the different agro-climatic zones within Europe. The results from this study provided us with an in-depth understanding of the factors influencing the uptake and use of DSS in IPM in different parts of Europe, and can be upgraded into extended research on how to improve the adoption and use of DSS in IPM in specific regions of Europe.

2. Materials and methods

Supporting the implementation of the Directive on the Sustainable Use of Pesticides (Directive 2009/128/EC) throughout the EU, we conducted a survey in 12 European countries to capture the variability of agro-environmental zones and the differences in economic and social development across European agriculture.

2.1. Survey

Primary data were obtained by surveying potential IPM DSS users. In 12 participating countries, institutions and companies involved in farm advisory services as well as farmers involved in the crop production sector of agriculture including farmers with mixed farms (e.g. livestock and crop production) were invited by email to their respective national IPM Decisions workshop. To ensure that participant representation was randomly distributed within target groups, all invitees were asked to invite other farmers or farm advisors, and the workshops were promoted through existing relevant national networks and social media. To avoid possible sampling bias, the workshops took place independently and were not part of a larger event. Participation was voluntary, free of charge, and participants were not compensated for answering our questionnaire. No further selection was made among the participants and all participants responded to the questionnaire. The face-to-face workshops were held in January

Table 1. Division of the countries participating in the study into three regions and the number of respondents for each region.

Region	Farmers	Farm advisors	Participating countries
Northern Europe	32	36	Denmark, Finland, Lithuania, Sweden, Norway - advisors only
Central Europe	32	46	Germany, Netherlands, Slovenia, UK
Southern Europe	36	37	Greece, Italy, France
Total	100	119	219 participants from 12 countries

and February 2020 (Table 1), and conducted in the national language of each country. Each participating country held one or two workshops in the location of the organising partner's choice. Separate questionnaires were developed for each target group ((i) farmers and (ii) farm advisors), and were consistent for all participating countries. Each participant identified him/herself as a farmer or a farm advisor and voluntarily completed the corresponding questionnaire. The questionnaires were printed on paper, and self-completed by the farmer/advisor; the workshop organisers were present in case clarification was needed. Both questionnaires were developed in the English language and reviewed by workshop organisers. It was then checked by a native English speaker and reviewed by the IPM Decisions project's Ethics Committee, before being translated into all 12 languages of the participating countries. The survey contained both direct and indirect, closed-ended questions. The questionnaires were designed based on the approaches, findings and suggestions for improvement from published research on similar topics (Norton et al. 1999; Gent, De Wolf, and Pethybridge 2011; Hochman and Carberry 2011; Aldhmour and Eleyan 2012; Shtienberg 2013; Kabir and Rainis 2015; Rose et al. 2016; Rose et al. 2018; Rossi et al. 2019).

The questionnaire contained 23 questions for farmers (Appendix A) and 15 questions for farm advisors (Appendix B), that asked participants about their demographic characteristics (e.g. age, gender); the characteristics of their farms (or farms managed by farm advisors); their experiences, habits and expectations regarding IPM DSS use; their attitudes towards IPM DSS use; legal requirements; and technical and communication capabilities. The questionnaire for farm advisors also included a series of questions aimed at assessing the attitude of their clients (farmers) towards IPM DSS. In total, 100 farmers and 119 farm advisors completed the questionnaire (Table 1). The data from the participating countries were grouped according to broad environmental and agroclimatic regions, corresponding to the environmental stratification of Europe: i) Atlantic – Northern Europe; ii) Continental – Central, and; iii) Mediterranean - Southern (Metzger et al. 2005), which corresponds to the country of the participants (Table 1). For the purposes of our study, we will use the term region (Northern Europe, Central Europe and Southern Europe).

2.2. Analysis

Each survey question represented an attribute in our database with ordinal or nominal qualitative values representing the answers to the question asked. The attributes were recoded to take the small sample size into account and to reduce dispersion of data. To ensure the comparability of the numerical and nominal attributes in the dimensionality reduction step, the whole dataset was split into two subsets: (i) dataset containing attributes with numeric and ordinal categorical attributes and (ii) dataset containing nominal attributes. Given the relatively small number of completed surveys, the data were analysed in three steps, i) dimensionality reduction, ii) clustering analysis and iii) classification models and complemented with iv) statistical analysis.

2.2.1. Dimensionality reduction

In the first step, a dimensionality reduction method (McInnes, Healy, and Melville 2018) was applied to the questionnaire responses data. By answering the questionnaire, each survey respondent (farmer or farm advisor) developed their own multidimensional data profile, which was described by 23 dimensions (survey questions/attributes) for farmers and 15 dimensions for farm advisors. Each dimension had several possible values (between 2 and 5) illustrating the general attitude towards IPM DSS and their use. The differences between respondents' answers are therefore differences in the values of each dimension. The information scattered in the multidimensional data profile (defined by the number of the questions), of each participant was integrated into two dimensions (x, y). In this way, the profile characteristics and relationships were preserved as much as possible. The distances between respondents in the higher-dimensional space remained similar to those in the reduced two-dimensional space, so that respondents with similar survey answers were closer together in the x, y plane and vice versa. This allowed us to visualise the distribution of survey respondents on the x, y plot.

Dimensionality reduction algorithms are unsupervised machine learning methods. For our purposes, we used Uniform Manifold Approximation and Projection (UMAP), (McInnes, Healy, and Melville 2018). This is a general-purpose manifold learning and dimensionality reduction algorithm that arguably preserves more of the global data structure by finding a low-dimensional projection or embedding of the data that has the best possible equivalent fuzzy topological structure (McInnes, Healy, and Melville 2018). Dimensionality reduction was performed separately for both datasets (i.e. numeric/ordinal and nominal attributes) using appropriate distance metrics: Euclidean for numeric/ordinal and Hamming for binary attributes (nominally coded),

Table 2. Settings of the UMAP and HDBSCAN algorithm parameters used for dimensionality reduction and clustering of farmers' and farm advisors' datasets.

	Farmers	Farm advisors
UMAP (ordinal attributes)	n_neighbors = 10, min_dist = 0.1, n_components = 2, metric = ″euclidean″, random_state = 123	n_neighbors = 5, min_dist = 0.5, n_components = 2, metric = ″euclidean″, random_state = 123
UMAP (nominal attributes)	n_neighbors = 10, min_dist = 0.1, n_components_components.1, n_componenrandom_state = 123	n_neighbors = 10, min_dist = 0.9, n_components = 2, metric = ″hamming″, random_state = 123
HDBSCAN	min_cluster_size = 5, min_samples = 4	min_cluster_size = 5, min_samples = 4

(Table 2). Before proceeding with clustering, the obtained low-dimensional embeddings (UMAP models) of the two datasets were combined using multiplication (intersection).

The dimensionality reduction and clustering was performed in Python within JupyterLab, the latest interface of the Jupyter project (Kluyver et al. 2016), using the scikit-learn machine learning library (Pedregosa et al. 2011), Pandas (McKinney 2010) and for visualisations seaborn (Waskom 2021) and Matplotlib (Hunter 2007). The same approach was used for clustering.

2.2.2. Clustering analysis

We used a density-based clustering method to objectively test for the existence of clusters, as the participants' integrated response profiles showed a tendency to cluster in the previous step. The HDBSCAN algorithm (Campello, Moulavi, and Sander 2013) was used to cluster survey respondents in the x, y plane and to automatically determine the most appropriate number of clusters. HDBSCAN can refuse to cluster some points and classify them as "noise" if they do not fit into any cluster, which allows the algorithm to "not be wrong". Since this is unsupervised learning, where there is no ground truthing to measure the classification error, the clusters obtained were evaluated using density-based clustering validation (DBCV), (Moulavi et al. 2014), which gives an indication of how well the clusters are separated from each other. DBCV is specifically suited for density-based clustering algorithms because it takes noise into account and captures the shape property of clusters using densities rather than distances. DBCV calculates two values, (i) the density within a cluster (cluster cohesion) and (ii) the density between clusters (cluster separation). A high density within a cluster and a low density between clusters indicates high clustering quality. The final result of DBCV is a weighted sum of the "Validity Index" values of clusters with a score between −1 and 1, where the better the clustering assessed, the greater the value.

2.2.3. Classification models

In order to identify the factors that differentiate the clusters between each other, in the final step of the analysis a machine learning method was used to generate classification trees (Frank, Hall, and Witten 2016). This allowed expansion of the two dimensional information back to the underlying dimensions, i.e. the attributes and their values.

The classification tree predicts the nominal value of the dependent variable. In this study, the dependent variable was the cluster of participants, and its nominal values are the values of the clusters discovered through cluster analysis. The classification tree (Breiman et al. 2017) is a hierarchical structure where each internal node contains a test for an attribute, each branch corresponds to an outcome of the test, and each leaf node (terminal node) provides a prediction for the value of the class variable (the number of the cluster). The interpretation of the classification trees follows the rule "if-then" from the top attribute (most influential attribute) in the classification tree downwards to the leaves representing the values of the dependent variable (in our case the cluster number), (Figure 1). The prediction accuracy of a classification tree is the proportion of successful predictions over the entire set of instances, which measures the overall performance of the classifier (Witten et al. 2016). The most commonly used technique for estimating the predictive performance of a classifier on unseen data is cross-validation. In cross-validation, a fixed number of folds or partitions of the data are defined. Due to the small number of instances (fully completed questionnaires), leave-one-out cross-validation was used (Witten et al. 2016). This is a special type of cross-validation where the number of folds is equal to the number of instances in the dataset. The classification algorithm is thus applied once for each instance, using all other instances as the training set and the selected instance as a single-item test set. The accuracies obtained in the different iterations are averaged to obtain an overall accuracy. To construct a classification decision tree, the algorithm J4.8 was used; an implementation of the C4.5 algorithm (Quinlan 1993) in the software package WEKA (Frank, Hall, and Witten 2016). The parameters of algorithm J4.8 were set to their default values. For unbalanced data (the farmers in Central Europe and the farm

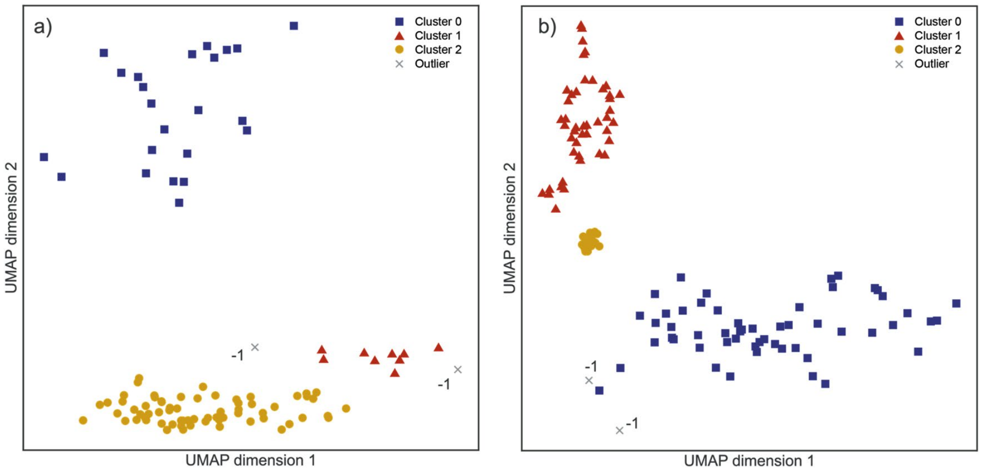


Figure 1. Arrangement of (a) farmers and (b) farm advisors in a two-dimensional plane and their grouping in three clusters (0, 1 and 2).

advisors in Northern Europe), the Cost Sensitive Classifier with reweighted training instances was selected as the classification method. The validation of the constructed model (i.e. the predictive performance of the classification) was assessed by the percentage of correctly classified instances.

To clarify the criteria for forming the clusters, we used the baseline datasets extended by the attribute "Cluster". Each participant was assigned a cluster number, which was determined during the cluster analysis. Only the most informative attributes (survey questions) were included in the hierarchical structure of the classification model, based on which the participants were classified into the discovered clusters. At the same time, these attributes also explain the variability between the clusters. The attributes incorporated into the structure of the classification model by the machine learning algorithm represent the explanatory criteria for the differences between the groups (clusters) of participants in their attitudes and opinions about IPM DSS.

2.2.4. Statistical analysis

Since the structure of the classification trees consists of only the most influential attributes, the cluster and classification analysis were additionally supplemented with a z-test to statistically analyse the differences between clusters. The importance of all attributes was thus further tested by a statistical analysis of the differences in the answers to the same question ($p < 0.05$; Table 4 and 5). This additional step identified other important but less influential attributes that explain the differences between the participants classified in different clusters. To avoid bias due to the small data sample and to support the identification and interpretation of the discovered patterns through data mining, a z-test for comparing two proportions (Equation (1)) was chosen and analysed with a 95% confidence interval (Zou et al. 2003).

$$z = \frac{p_1 - p_2}{\sqrt{\left[p_c\left(1-p_c\right)\right] / \left(1/n_1 + 1/n_2\right)}}$$

(1)

Where:

$$p_c = \frac{Y_1 + Y_2}{n_1 + n_2}$$

(2)

p_1 = proportion of answer "A" in group 1
p_2 = proportion of answer "A" in group 2
p_c = proportion of answer "A" in the sample
Y_1 = number of participants who answered "A" in group 1
Y_2 = number of participants who answered "A" in group 2
n_1 = sample size of group 1
n_2 = sample size of group 2
"A" = attribute value (e.g. Yes; No)

3. Results

The age structure of farmers and farm advisors was similar in all regions, with a slightly higher proportion of younger farmers (25–35 years old) participating in Southern Europe. Of the farmers, 27% were 25–35 years old, 28% were 35–50 years old and 27% were 50–60 years old. There were few younger (10% aged

Table 3. The results of the DBCV validation of clustering with HDBSCAN for farmers and farm advisors in all three regions of Europe and all Europe together.

	Northern Europe		Central Europe		Southern Europe		All Europe	
Participants	Farmers	Advisors	Farmers	Advisors	Farmers	Advisors	Farmers	Advisors
No. of participants	32	36	32	46	36	37	100	119
No. of attributes	23	15	23	15	23	15	23	15
No. of HDBSCAN clusters	2	2	2	2	2	3	3	3
DBCV score	−0.18	0.25	−0.26	0.63	−0.09	0.36	0.25	−0.12

Table 4. Statistically significant differences ($p < 0.05$) between the responses of farmers from the two clusters (0 and 1) in all three regions of Europe.

Attribute	Northern Europe	Central Europe	Southern Europe
Already used DSS	* ♦	*	*
Need IT teaching	* ♦	*	*
Trust in DSS	* ♦	*	*
Productivity improvement	*		*
Appreciate change		*	
Enjoy using new technology		* ♦	
Farm size		*	
Gender		*	
Importance of low price		*	
Importance of easy to use		*	
Income		*	
Legislative requirements		*	
WTP for DSS		*	
DSS must be tailored			*
Importance of efficiency			*
Importance of polyvalence			*
Followed manufacturer's demo			*
High-speed internet			* ♦
Willing to try new products			* ♦

"*"indicates attributes for which we found statistically significant differences between the clusters, while "♦" indicates attributes that were included in the classification tree in each region.

18–25) and older (8% aged > 60) farmers. The majority (61%) of farm advisors surveyed were between 35 and 60 years old, 29% were between 18 and 35 years old and 11% were over 60 years old. The vast majority (87%) of farmers were male, while the gender structure is more balanced among farm advisors, with 57% males and 43% females. Of the farmers, 75% had at least a bachelor's degree, while farm advisors with a vocational degree are underrepresented in all three regions, due to the professional requirements for higher education. Farmers with small (less than 5 ha) farms (most in Central Europe (22%)), and farm advisors who predominantly advise small farms (9%) or only large farms (7%) are underrepresented in all three regions. The majority of farmers were from large (> 60 ha) farms, and farm advisors in all three European regions (47%) advise farms of mixed size, followed by medium-sized farms (37%). The under-representation of certain groups of farmers and farm advisors (e.g. farmers with smaller farms (<5 ha)) was the result of the decision of those invited to participate or not. A more detailed description of the data categories mentioned in Section 2.1. can be found in the following subsections.

Table 5. Statistically significant differences ($p < 0.05$) between the responses of farm advisors from both clusters (0 and 1) in Northern and Central Europe and all three clusters (0, 1 and 2) in Southern Europe. "*" indicates the attributes for which we found statistically significant differences in each region, while the symbol "♦" indicates the attributes that were included in the classification model in each region.

Attribute	Northern Europe	Central Europe	Southern Europe
Access to marketing info about DSS	* ♦	*	*
Followed manufacturer's demo	*	* ♦	
Age		*	* ♦
Productivity improvement	*		
High-speed internet	*		
Experience		* ♦	
Output preference		*	
Crop type group			* ♦

3.1. Data clustering

The results of the cluster analysis provide insight into the variability of survey respondents' answers regarding their experiences and expectations of using DSS in IPM. Participants who gave the same answer to most questions as other participants were distributed closer together on the x, y plot. Both farmers (Figure 1a) and farm advisors (Figure 1b) show tendency for forming into three groups. Outliers were found in both farmers and farm advisors, that were not classified into any of the discovered clusters by the HDBSCAN algorithm. The discovered clusters in the data of farmers from all participating countries reached a DBCV score of 0.25, while data of farm advisors in the same countries reached a DBCV score of −0.12. (Table 3).

The structure of the classification tree showed that the clusters, confirmed by the results of the cluster analysis shown in Figure 1, are mostly influenced by the region from which the participants come. Due to variability of the responses, the group of farmers in Southern Europe was further divided in the classification tree into a group of participants who have already used an IPM DSS and a group of participants who have not yet used an IPM DSS. The accuracy of the classification tree, determined by leave-one-out cross-validation (Witten et al. 2016) for farmers is 96.0% and 94.7% for farm advisors.

3.2. Data clustering by region

The results of the initial data clustering analysis suggest that the data can be divided by region, therefore, the analysis was repeated using data divided by region for farmers and for farm advisors. Participants were clustered into two or three identified clusters (Table 3).

The results of the clustering method (HDBSCAN) of data from specific region indicated a tendency to form clusters of farmers' and farm advisors' data in all three regions of Europe. Visualisation of HDBSCAN of two groups show that farmers (Figure 2a) in Southern Europe were clustered in two clusters while farm advisors (Figure 2b) in Southern Europe were clustered in three clusters.

3.3. Influential attributes of discovered clusters by region

The results of the classification trees constructed using the data from each region show differences between the regions. Attributes with statistically significant differences among both farmers and farm advisors are described in separate subsections for each region. Statistically significant differences ($p < 0.05$) were found between the responses of farmers classified in two clusters (Table 4). All attributes included in the farmers' analysis and their values (possible answers) are listed and described in Appendix A.

Statistically significant differences ($p < 0.05$) were found between the responses of farm advisors classified in both or all three clusters (Table 5). All the farm advisors' attributes included in the analysis and their values (possible answers) are listed and described in Appendix B.

Based on the structure of the classification model and the results of the statistical analysis, we determine which group(s) of farmers or farm advisors are more inclined to use IPM DSS.

3.3.1. Farmers in Northern Europe

Farmers in Northern Europe were classified into two clusters (0 and 1). In the structure of the classification tree, the attribute "Already used DSS", which describes whether farmers have already used DSS, played the most important role, followed by the attributes "Trust in DSS" and "Need IT Teaching" (Figure 3; Table 4), which describe their trust in DSS and their perceived need for additional education and training in IT.

3.3.1.1. Northern Europe farmer cluster 0 – more inclined to use IPM DSS. These were farmers who already used IPM DSS and moderately or strongly trusted DSS, and those who already used IPM DSS but do not trust them and feel that they do not need additional information technology (IT) education. Posthoc analysis showed that farmers in this group

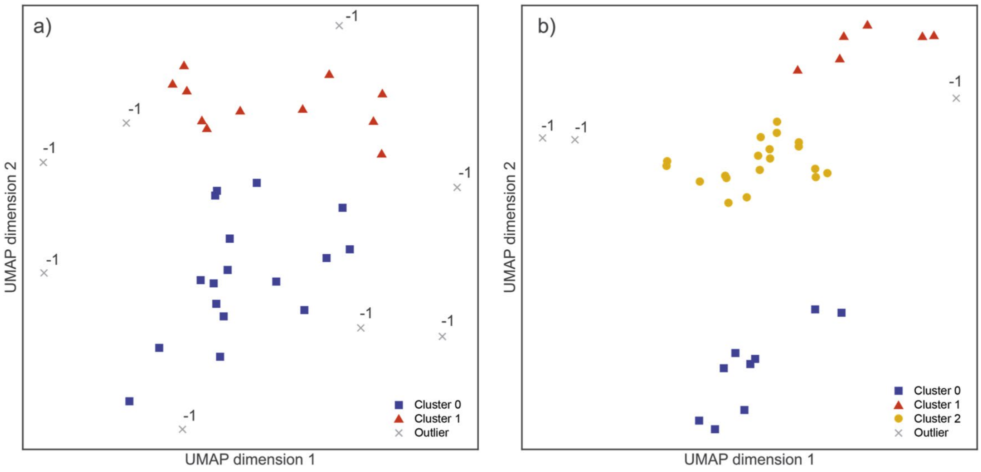


Figure 2. Arrangement of (a) farmers and (b) farm advisors in Southern Europe in a two-dimensional plane and their grouping in two (farmers) or three (farm advisors) clusters.

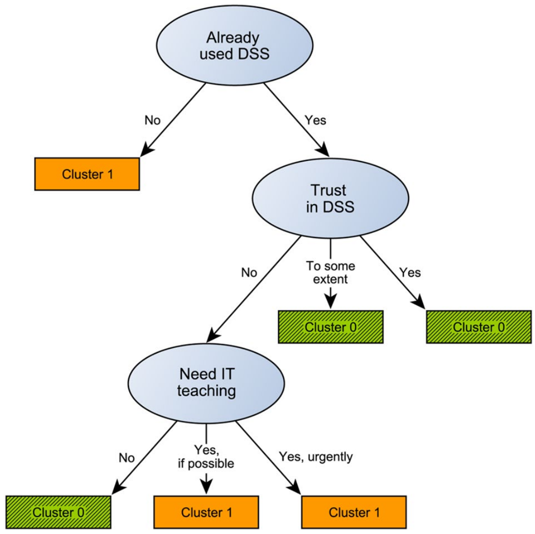


Figure 3. Classification model for farmers in Northern Europe. The attribute values of farmers classified in cluster 0 are striped and highlighted in green, while the attribute values of farmers classified in cluster 1 are highlighted in orange. The model correctly classified 78.1% of the instances.

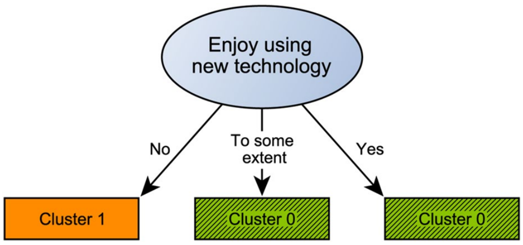


Figure 4. Classification model for farmers in Central Europe. The attribute value of farmers classified in cluster 0 are striped and highlighted in green, while the attribute values of farmers classified in cluster 1 are highlighted in orange. The model correctly classified 81.3% of the instances.

that were more inclined to use IPM DSS were more likely to think that IPM DSS can have a positive impact on their productivity and decision-making ability (“Productivity improvement”) than farmers that were assumed to be less inclined to use IPM DSS ($p < 0.05$; Table 4). Therefore, farmers classified in cluster 0 could be described as “(potential) users of IPM DSS”.

3.3.1.2. Northern Europe farmer cluster 1 – less inclined to use IPM DSS. These were farmers who had not yet used IPM DSS or those who had already used IPM DSS but do not trust them (“Trust in DSS”) and at the same time admitted that they need further training in IT (“Need IT teaching”), ($p < 0.05$; Table 4). Posthoc analysis showed that farmers in this group were less inclined to use IPM DSS and could therefore be described as “sceptics towards the use of DSS”.

3.3.2. Farmers in Central Europe

Farmers in Central Europe were classified into two clusters (0 and 1). The classification tree has only one linkage (Figure 4). This is due to the small number of participants in the region (Table 3). Clusters were differentiated based on the attribute “Enjoy using new technology”, which describes farmers’ willingness to use new technologies. Farmers who enjoy using new technologies to some extent or very much are placed in cluster 0 and farmers who do not enjoy the use of new technologies are placed in cluster 1.

3.3.2.1. Central Europe farmer cluster 0 - more inclined to use IPM DSS. Farmers in cluster 0 were more inclined to use new technologies compared to farmers classified in cluster 1, and more cluster 0 farmers were already using DSS (90%, “Already used DSS"; $p < 0.05$; Table 4). Of the farmers in this cluster, 46% responded that they do not need additional IT training, and since farmers in this cluster were associated with larger farms, they could therefore be described as “farmers with larger farms and better IT knowledge”. Posthoc analysis ($p < 0.05$; Table 4) showed that majority of farmers classified in cluster 0 trust DSS (“Trust in the DSS”), (86%), are more open to change (“Appreciate change”), (73%), have larger farms (“Farm size”), (73%), higher income from farming (“Income”), and have to a lesser extent indicated that feel they need additional IT training (“Need IT teaching”), compared to the farmers classified in cluster 1 (assumed to be less inclined to use DSS). The importance of ease of use is of medium to high importance for this cluster (“Importance of easy to use”), and 95% of farmers in this cluster reported market or legislative requirements regarding the use of pesticides and fertilisers (“Legislative requirements”). Farmers who were assumed to be more inclined to use DSS give less importance to the low price (“Importance of low price”) and are more willing to pay for DSS (“WTP for DSS”) even if no subsidies are available. The majority of participants in cluster 0 were male (“Gender"; 91/9).

3.3.2.2. Central Europe farmer cluster 1 – less inclined to use IPM DSS. These farmers could be described as “farmers with smaller farms who are averse to using new technologies”, and relatively fewer were already using DSS (60%) compared to in cluster 0. The importance of ease of use is of high importance

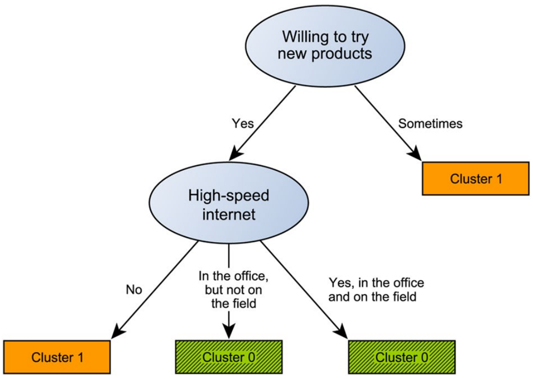


Figure 5. Classification model for farmers in Southern Europe. The attribute values of farmers classified in cluster 0 are striped and highlighted in green, while the attribute values of farmers classified in cluster 1 are highlighted in orange. The model correctly classified 80.6% of the instances.

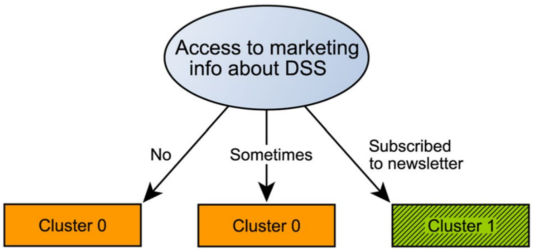


Figure 6. Classification model for farm advisors in Northern Europe. The attribute value of farm advisors classified in cluster 0 are highlighted in orange, while the attribute value of farm advisors classified in cluster 1 are striped and highlighted in green. The model correctly classified 57.1% of the instances.

for farmers who we assume are less inclined to use IPM DSS, and 60% reported market or legislative requirements on the use of pesticides and fertilisers. They were also less willing to pay for DSS ("WTP for DSS"). The gender distribution ("Gender") of participants in this cluster was more evenly balanced (60/40) than in cluster 0.

3.3.3. Farmers in Southern Europe

Farmers in Southern Europe were also classified into two clusters (0 and 1) using the clustering method. In the structure of the classification tree (Figure 5), the attribute "Willing to try new products", which describes the willingness to try new technologies and services, played the most important role, followed by the possibility to have access to high-speed internet ("High-speed internet").

Cluster 0 included farmers who are willing to use new products and services but have access to the internet only in the office (at home) and not on all their fields, and farmers who have access to high-speed internet both at home and on all their fields. Cluster 1 included farmers who are slightly more reluctant to use new products or farmers who like to use new products and services but have an infrastructural barrier of not having access to high-speed internet.

3.3.3.1. Southern Europe farmer cluster 0 - more inclined to use IPM DSS. Farmers classified in cluster 0 are "farmers who like to use new technologies", though 96% of this group responded that they need additional IT training ("Need IT teaching"). The majority of these farmers occasionally attend DSS demonstration workshops ("Followed manufacturer's demo"), (70%) and had a much better attitude towards DSS as more than 95% of them are convinced that DSS can help them in decision-making and productivity ("Productivity improvement"). These farmers trust DSS to a greater extent ("Trust in DSS") and a higher proportion have already used them ("Already used DSS"), (75%). Farmers more inclined to use IPM DSS consider the polyvalence of DSS ("Importance of polyvalence"), i.e. property of a DSS to predict multiple pests or pests for multiple crops, to be important ($p < 0.05$; Table 4). Farmers from cluster 0 expressed the wish that decision support systems should be tailored to their farm ("DSS must be tailored"), ($p < 0.05$; Table 4).

3.3.3.2. Southern Europe farmer cluster 1 – less inclined to use IPM DSS. Farmers classified in cluster 1 mostly answered that they have never used a DSS and could be referred to be "farmers with lack of experience in using a DSS". Most of these farmers responded that they needed IT training (71%), 70% had never attend DSS demonstration workshops ("Followed manufacturer's demo"), and almost 40% responded that DSS cannot help them in decision making and productivity ("Productivity improvement"). Of these farmers, 57% have not yet tried DSS ("Already used DSS"), though considered it extremely important that DSS are efficient ("Importance of efficient"), ($p < 0.05$; Table 4). Farmers in cluster 1 also wanted decision support systems to be tailored to their farm.

3.3.4. Farm advisors in Northern Europe

The farm advisors in Northern Europe were classified into two clusters (0 and 1) using the clustering method. The decision tree has only one link (Figure 6). This is due to the small number of instances (Table 3). The algorithm chose "Access to marketing info about DSS", describing availability of information provided by DSS developers (i.e. name of the DSS, the problem it addresses, access link, price etc.), as the most discriminating attribute between clusters 0 and 1. Cluster 0 consisted of farm advisors who do not have access to marketing information about DSS or those who only have occasional access to marketing information about DSS. Cluster 1 consisted of farm advisors who subscribe to news about DSS.

3.3.4.1. Northern Europe farm advisor cluster 0 – less inclined to use IPM DSS. These farm advisors have relatively poorer access to information about DSS, as 48% do not have access to market information on DSS and 52% only have limited access to market information on DSS. Most of these farm advisors reported a partial improvement in farm productivity when using DSS ("Productivity improvement"), (61%), 30% reported a great improvement and 9% either did not notice any improvement or have not yet used DSS. Most occasionally attend manufacturer's demonstration workshops about DSS when invited ("Followed manufacturer's demo"), (61%), and others did not attend them (35%). The majority of farm advisors in this cluster have access to high-speed internet in the office and in their clients' fields (61%), but the proportion of advisors who do not have access to high speed internet in clients' fields ("High-speed internet") is quite high (35%***).***

3.3.4.2. Northern Europe farm advisor cluster 1 – more inclined to use IPM DSS. These farm advisors had relatively good access to information about DSS, i.e. "subscribers to market news on DSS" ("Access to marketing info about DSS"). Most of these farm advisors had positive experience wit***h*** DSS: 69% reported a great improvement in productivity using DSS and the rest (31%) reported a partial improvement through DSS ("Productivity improvement"). Most (69%) occasionally attend manufacturer's demonstration workshops about DSS when invited ("Followed manufacturer's demo"), and the rest (31%) attended them regularly. Of these farm advisors, 92% had access to high-speed internet both in the office and in clients' fields (High-speed internet").

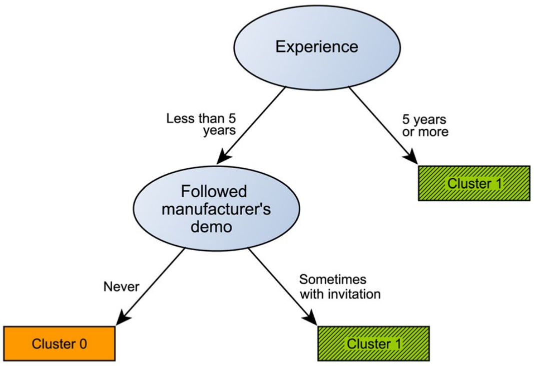


Figure 7. Classification model for farm advisors in Central Europe. The attribute value of farmers classified in cluster 0 are highlighted in orange, while the attribute value of farm advisors classified in cluster 1 are striped and highlighted in green. The model correctly classified 100% of the instances.

3.3.5. Farm advisors in Central Europe

The farm advisors in Central Europe were classified into two clusters (0 and 1) using the clustering method. In the structure of the classification tree (Figure 7), the attribute ("Experience") played the most important role. Cluster 0 consisted of farm advisors with less than 5 years of experience in agricultural extension who had never attended a DSS demonstration ("Followed manufacturer's demo"). Cluster 1 consisted of farm advisors with less than 5 years' experience in agricultural extension who occasionally attend DSS demonstration workshops, and those with more than 5 years' experience in farm extension.

3.3.5.1. Central Europe farm advisor cluster 0 – less inclined to use IPM DSS. All these farm advisors had less experience in agricultural extension (< 5 years), ("Experience"), and do not have access to marketing information about DSS ("Access to marketing info about DSS") and the majority were younger (18 - 35 years old), (83%). Of the farm advisors in this cluster, 67% had a higher preference for simple graphical presentation (lower complexity) of DSS results ("Output preference"***).***

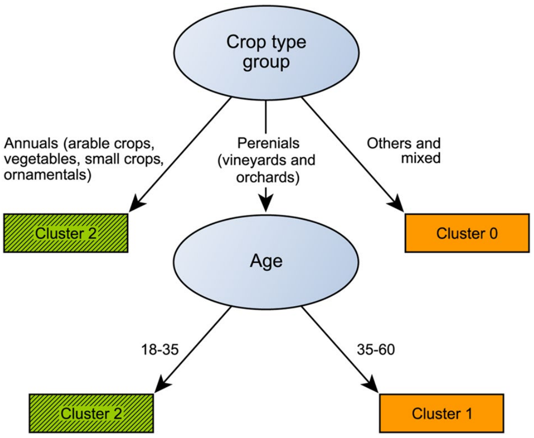


Figure 8. Classification model for farm advisors in Southern Europe. The attribute values of farmers classified in clusters 0 and 1 are highlighted in orange while the attribute values of farm advisors classified in cluster 2 are striped and highlighted in green. The model correctly classified 94.1% of the instances.

3.3.5.2. Central Europe farm advisor cluster 1 – more inclined to use IPM DSS. The majority of these farm advisors had more than 5 years' experience in agricultural extension ("Experience"), (88%). Compared to advisors in cluster 0, the majority is relatively older (35-50 years old), (73%). Most have occasional access to marketing information about DSS (70%). Most farm advisors in this cluster prefer more detail in DSS outputs ("Output preference"), (80%). They felt a simple graphical presentation of the results is not sufficient and they would like an additional explanation in text form.

3.3.6. Farm advisors in Southern Europe

The farm advisors in Southern Europe were classified into three clusters (0, 1 and 2) using the clustering method, as their responses were more diverse compared to other groups of farmers and farm advisors. In the structure of the classification model (Figure 8), the attribute "Crop type group", which describes the group of crops predominantly grown by their clients (farmers), played the most important role. This was followed by the age of the farm advisors ("Age"). Cluster 0 consisted of farm advisors advising mixed farms and farms growing other crops not covered by the previous categories. Cluster 1 consisted of farm advisors who mainly advise farms growing perennial crops (orchards and vineyards) and are between 35 and 60 years old. Cluster 2 consisted of farm advisors who advise farms growing annual crops (arable crops, vegetable crops, minor crops or ornamental plants) and farm advisors who mainly advise farms growing perennial crops (orchards and vineyards) and are between 18 and 35 years old.

3.3.6.1. Southern Europe farm advisor cluster 0 – less inclined to use IPM DSS. These farm advisors mainly advise mixed farms and could therefore be referred to as "farm advisors for mixed farms". In this cluster, 67% reported no access to marketing information about DSS, and the remaining 33% had only occasional access.

3.3.6.2. Southern Europe farm advisor cluster 1 – less inclined to use IPM DSS. These farm advisors advise farms that grow grapes and fruit and are between 35 and 60 years old. Most of them (67%) have occasional access to marketing information about DSS. However, a large proportion of them (33%) do not have access and none of them are subscribers to marketing information about DSS (Table 6).

3.3.6.3. Southern Europe farm advisor cluster 2 – more inclined to use IPM DSS. These farm advisors could be described as "advisors for farms producing arable crops, vegetables and ornamental plants, and younger generation advisors for viticulture and orchards". Most of these advisors (64%) have occasional access to market information about DSS, 27% have no access, and 9% subscribe to marketing information (Table 6).

3.3.7 Barriers to IPM DSS uptake in Europe

Based on participant's responses to the questionnaire, we assumed each discovered cluster as "more" or "less" inclined to use IPM DSS, compared to other cluster in the same region and the same user type (farmer or farm advisor). We extracted the key characteristics of farmers and farm advisors who were labelled as "less inclined" to use IPM DSS and compared their characteristics with those of farmers and farm advisors, who were labelled "more inclined" to use IPM DSS, and identified some of these characteristics as barriers to IPM DSS adoption and use. These barriers are summarised in Table 7.

4. Discussion

Due to the many challenges in improving the uptake and use of IPM DSS among decision makers, our research focused on identifying the attributes among farmers and their advisors that are associated with uptake and use of DSS in IPM at regional and European levels. To address this challenge, we conducted a survey among farmers and farm advisors in 12 European countries. Among farm advisors in all three regions of Europe, we found a high level of understanding of the purpose of IPM DSS. Most farm advisors answered that IPM DSS are a complement to, and not a substitute for, their work. IPM DSS outputs are often based on one or more assumptions about the crop or previous management and need to be critically

Table 6. Statistically significant differences ($p < 0.05$) between the responses of farm advisors in Southern Europe for each cluster (0, 1 and 2).

Difference between clusters ($p < 0.05$)	Crop type group	Access to marketing info about DSS	Age
cluster 0 – cluster 1	*		
cluster 0 – cluster 2	*	*	
cluster 1 – cluster 2	*		*

"*" indicate attributes where statistically significant differences were found between the two clusters.

Table 7. Identified barriers to uptake of IPM DSS among farmers and farm advisors in Europe that were labelled as less inclined to use IPM DSS based on their responses to the questionnaire.

	Barriers to uptake of IPM DSS by Farmers	Barriers to uptake of IPM DSS by Farm advisors
Northern Europe	• Lack of experience with DSS. • Lack of trust in DSS. • Perceived need for training in IT.	• Low participation in DSS demonstration workshops. • Many only have internet access in the office, but not in their clients' fields. • Poor access to marketing information about DSS.
Central Europe	• Lack of trust in DSS. • Less open to using new technologies and to changes in general. • Low willingness to pay for DSS, and low price of DSS is preferred. • Perceived complexity of DSS; easy to use systems are preferred. • Perceived need for training in IT.	• Low participation in DSS demonstration workshops. • Poor access to marketing information about DSS.
Southern Europe	• Lack of experience with DSS. • Lack of trust in DSS. • Low participation in DSS demonstration workshops. • Low expectation for DSS to improve productivity. • Poor access to high-speed internet. • Perceived need for training in IT.	• Crop specific requirements and expectations from DSS. Special attention needs to be paid to crop types. • Poor access to marketing information about DSS.

evaluated before they are implemented. Farm advisors therefore play an important role in promoting and increasing the use of DSS among farmers (Rose et al. 2016). Communication between the farm advisors and farmers is crucial to successfully increase the use of DSS in agriculture, as it provides farmers with information on the reliability of DSS from someone they trust (Kuehne, Nettle, and Llewellyn 2019).

4.1. Farmer attributes associated with uptake of IPM DSS across Europe

Farmers less inclined to use IPM DSS in all three regions of Europe share common barriers to the use of IPM DSS, namely a lack of trust in DSS and the perception that they lack IT knowledge.

4.1.1. Farmers lack of trust and lack of experience with DSS

User trust in the system is crucial for the successful use of DSS (Lawrence, Goodwin, and Fildes 2002; Gent, De Wolf, and Pethybridge 2011; Rose et al. 2016). Trust is built on exposure to, and personal experience with, DSS. We found that a lack of trust in IPM DSS is a major barrier to uptake among farmers in Northern, Central, and Southern Europe which is in line with other research on this topic (Gent, De Wolf, and Pethybridge 2011; Rose et al. 2016). Many farmers, especially in Northern and Southern Europe, also had limited personal experience with DSS and had never attended a DSS demonstration. Meanwhile, farmers already using DSS, and/or attending demonstrations, had a higher level of trust in DSS. Greater access to IPM DSS demonstration workshops has been shown to improve adoption of improved farming approaches, using peer-to-peer learning approaches (Cooreman et al. 2018; Ingram et al. 2018). Farmers can witness and discuss the application of IPM DSS in practice, providing an opportunity for participants to share positive experiences with IPM DSS use and become their ambassadors. Demonstration workshops have a key role in convincing participants of the usefulness of DSS, addressing the low expectations for DSS to improve their productivity that has been identified as a barrier to uptake of IPM DSS. Showcasing the impact of DSS as part of a wider IPM strategy, and providing quantitative evidence of the costs and benefits, as well as the context specific considerations, is critical to improving uptake of DSS.

4.1.2. Farmer's experienced ease of use of DSS and their perceived lack of IT knowledge

Ease of use of IPM DSS and a perceived lack of IT knowledge were identified as barriers to uptake in Europe, similar to Hochman and Carberry (2011) research. These can be overcome through training and/or demonstration workshops about DSS. Although most of the DSS developed for IPM are not very complex or difficult to use, it still takes some time to become familiar with them and use them independently. Farmers attending training on the use of DSS will need less time to familiarise themselves with using the DSS and are more likely to adopt further DSS in the future. Such training workshops should focus on how to use DSS and interpret outputs, and should aim to eliminate the perception among potential users that they do not have sufficient IT skills to use IPM DSS.

4.1.3. Farmer's willingness to pay for IPM DSS

Farmers in Central Europe could be grouped by their willingness to pay for DSS. The first group consists of farmers with smaller farms, of which a very high proportion (40%) were not willing to pay for DSS at all, and an equally high proportion (40%) were only willing to pay if there is a subsidy for the use of such systems. In contrast, the vast majority of farmers in the second group are willing to pay even if no subsidy is available. This is an important barrier to DSS adoption among farmers, and the introduction of subsidies for the purchase of or subscription to DSS at both national and European level is likely to increase uptake. In Southern Europe, there was more widespread reluctance to pay for IPM DSS.

4.1.4. Farmer's access to high-speed internet

A very important barriers for farmers in Southern Europe is access to high-speed internet, which can be a major infrastructural barrier. This particularly impacts DSS intended for use in the field that require internet connection for operation and data entry. A large proportion of farmers who were assumed to be less likely to use a DSS based on their responses to the questionnaire did not have access to high-speed internet across their farmland.

4.2. Farm advisor attributes associated with uptake of IPM DSS across Europe

A common barrier to the use of DSS among farm advisors in all three regions of Europe is poor access to marketing information about DSS.

4.2.1. Farm advisor access to IPM DSS market information and IPM DSS demonstrations

Across Europe, access to market information on IPM DSS and participation in demonstration workshops on DSS were very important factors influencing farm advisor attitudes towards IPM DSS. Farm advisors who have better access to market information on IPM DSS also have more positive attitudes towards DSS use. Access to market information about DSS was the biggest barrier to DSS adoption. This is in line with the findings of Rose et al. (2016), who described level of marketing as a driving factor for the adoption and use of DSS. This barrier could be systematically addressed through regionally coordinated training programmes, and access to demonstration workshops.

4.2.2. Farm advisor access to high-speed internet

Most farm advisors in Northern Europe have access to high-speed internet, both in the office and on all their clients' farmland. However, a large proportion (35%) of farm advisors in Northern Europe, whom we assumed to be less likely to use DSS based on their responses, have access to high-speed internet only in the office, but not on all their clients' fields. The infrastructural barriers already described in the case of farmers in Southern Europe were also found among farm advisors in Southern Europe, as a large proportion of farm advisors in Southern Europe answered that they only have access to high-speed internet at home or in the office, but not in all areas of their clients' fields. Developers of DSS in this region are therefore advised to develop systems that are designed for office use, or those that allow data entry from the fields even without a high-speed internet connection.

4.2.3. Farm advisor crop specific uptake

In Southern Europe, the use or willingness to use DSS is related to the type of crop for which they advise. Among farm advisors, we found that their expectations and needs regarding DSS strongly depend on the type of crop grown on their clients' farms. As Rose et al. (2016) have already pointed out, the relevance of DSS is of great importance for the uptake and use of DSS. The results of our survey show that farm advisors in Southern Europe who advise farms where mainly annual crops (arable, vegetables and ornamentals) are grown are the most inclined to use IPM DSS.

4.3. Application of results

We are facing increasingly severe impacts of climate change (Shukla et al. 2019) and the accompanying increased pressure of some pests on some crops and increased use of some pesticides (Bale et al. 2002; Taylor et al. 2018). Increased uptake and access to IPM DSS would help to improve the decision-making process for a more optimal and efficient use of pesticides and consequently enable a reduction in pesticide use (Johnen and Meier 2000; Caffi, Rossi, and Bugiani 2010; Jones et al. 2010; Caffi et al. 2012; Rossi et al. 2014; Damos 2015, Kanatas et al. 2020; Lázaro, Makowski, and Vicent 2021), thereby pursuing the objectives of sustainable use of pesticides as defined in the Directive on the Sustainable Use of Pesticides (Directive 2009/128/EC). The complementary use of data mining and statistical methods allowed us to overcome the analytical limitations imposed by the structure and size of the data collected through questionnaires. This approach allowed us to identify the main barriers and incentives to DSS adoption among farmers and their advisors across Europe. We demonstrated that the data mining methods that are primarily applied to large data sets (Witten et al. 2016) can also be used to extract information from relatively small data sets.

Across Europe, the key barrier for farmers was a lack of experience and trust in IPM DSS. Underpinning this are concerns that either individual DSS do not adequately reflect the conditions on their farm, or that the DSS outputs do not fit with their existing crop protection strategy (Rossi, Caffi, and Salinari 2012). Increased transparency of IPM DSS development, and associated assumptions and limitations, is needed to engage and build trust between developers, farmers and farm advisors. In addition, regional demonstration of DSS in practice within a relatable context, will encourage farmers and their advisors to trial the use of IPM DSS on their farms. The most effective route to achieving this is through open access to a range of transparent IPM DSS, and regional networks demonstrating IPM DSS in action (Rossi et al. 2019). Improving support provided to farmers by their advisors will also support improved trust and uptake; however, we also found that the farm advisors needed access to improved market information about IPM DSS to increase uptake. Improved support can be achieved through the same regional demonstration networks, providing IPM DSS training and workshops in parallel to demonstration activities (Lacoste and Powles 2016). Over time, as novel IPM DSS are developed, crop management strategies evolve, and climate change leads to changes in pest ecology, farmers and their advisors will require ongoing access to up-to-date systems, approaches, and training, requiring long term support networks.

Education and training with IPM DSS could be extended in the curricula of agricultural high schools, colleges and universities in order to adapt to environmental standards and industry progress (Durmuş and Dağli 2017). Such education or training programmes could be longer than the one-day workshops mentioned above and could take place throughout the growing season, so that both young farmers and potential future farm advisors get a good introduction to the use of such tools (Voogt et al. 2005; Zhou et al. 2015). This requires a high degree of willingness on the part of professors or the involvement of professors in seeking external help (e.g. developers or providers of DSS). Since the developers or providers of DSS would engage with potential end users through this collaboration, we would expect them to be open to collaboration (Bruneel, d'Este, and Salter 2010). Improving stakeholder involvement in DSS development, either through demonstration networks or extended training programs, would also promote trust in their application (Lacoste and Powles 2016). Developers would get much needed feedback from potential users and develop tools that have a much higher utility value and could therefore increase the adoption of such DSS.

In addition to improving trust in IPM DSS, there is also a need for investment in essential infrastructure and access. In all parts of Europe, but most notably in the Southern region, a lack of high-speed internet in the field (and in some areas in the office) is a barrier to IPM DSS uptake, along with any direct costs associated with access to IPM DSS. This requires National and international government incentives and/or agri-environmental policy changes, including investment in infrastructure and/or introduction of subsidies to facilitate access to and use of DSS, alongside support for advanced technologies that contribute to reduced use of pesticides and more effective implementation of IPM practices. In practice, this requires financial and practical support for open access IPM DSS, demonstration networks, and training at a national level.

5. Conclusions

We have identified the barriers and incentives to access and uptake of DSS by farmers and farm advisors across Europe, as well as the specific barriers and incentives in the different regions of Europe. This will allow us to steer future research towards finding appropriate region-specific solutions to overcome these barriers. We have also proposed a method for analysing the data collected by questionnaires, which are characterised by a large amount of information scattered across the usually wide range of answers to the questions. Our methodological approach proved to be effective despite the relatively small number of fully completed questionnaires. We were able to extract important insights and make informed interpretations from the very scattered information provided by the answers to the questions.

We have identified the need to organise educational or demonstration workshops as well as including DSS training in agricultural schools and universities as one of the most promising pathways to increase the uptake of IPM DSS in Europe, which should aim at informing farmers and farm advisors about the features and benefits of IPM DSS. The organisation of such training programs should be incorporated into course at agricultural high schools, colleges, universities; or applied research organisations in cooperation with the developers or providers of IPM DSS; or by national or regional extension services. In addition, national support to improve access to IPM DSS through open access platforms, and improved high-speed internet access would encourage greater use of IPM DSS on farms across Europe.

With the implementation of the proposed measures to overcome the identified barriers to the adoption and use of DSS in IPM across Europe, we could expect a gradual increase in the use of DSS among farmers and farm advisors. Increasing implementation of IPM practices would contribute to a more sustainable use of pesticides and a reduced impact of pesticides on the environment, thus pursuing the objectives of sustainable use of pesticides as defined in the Directive on the Sustainable Use of Pesticides (Directive 2009/128/EC).

Although the research focused on identifying region-specific barriers to DSS adoption in Europe, the findings can also be applied in a broader context in other parts of the world. The lack of accessible information about IPM DSS, the low level of trust in IPM DSS, and the lack of accessible information infrastructure (e.g. no access to high-speed internet) should be taken into account in all countries and/or by all institutions trying to improve the adoption of IPM DSS among farmers and farm advisors.

Notes

1. Polyvalence in this context is the property of a DSS to predict multiple pests or pests for multiple crops.
2. Marketing information about DSS is information about the name of the DSS, the problem it solves, where (e.g. link) you can access it, where (e.g. country, region) you can use it (validation data), the data needed, its price, etc.

Acknowledgements

We would like to thank Prof. Dr. Nikolaos Georgantzis from the Burgundy School of Business for developing the questionnaire and PhD student Jotham Akaka from the Economics Department & Laboratorio de Economia Experimental, Universitat Jaume I for digitising the raw data collected with the questionnaires. We would also like to thank the facilitators of the workshops of the IPM decisions project who conducted workshops in 12 European countries and the participants of the workshops for their responses to the questionnaire. We would also like to thank Lenka Trdina of the Jožef Stefan Institute for her help with the graphic editing of figures in this article.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This study was conducted as part of the European project IPM Decisions. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement 817617) and is supported by the Slovenian Research Agency (grant P2-0103).

Data availability statement

Data available on request from the authors.

References

Aldhmour, F. M., and M. B. Eleyan. 2012. "Factors Influencing the Successful Adoption of Decision Support Systems: The Context of Aqaba Special Economic Zone Authority." *International Journal of Business and Management* 7 (2): 163. doi: 10.5539/ijbm.v7n2p163.

Ali, S., M. I. Ullah, A. Sajjad, Q. Shakeel, and A. Hussain. 2021. "Environmental and Health Effects of Pesticide Residues." In *Sustainable Agriculture Reviews 48: Pesticide Occurrence, Analysis and Remediation Vol. 2 Analysis*, 311–336. Cham: Springer International Publishing.

Ara, I., L. Turner, M. T. Harrison, M. Monjardino, P. DeVoil, and D. Rodriguez. 2021. "Application, Adoption and Opportunities for Improving Decision Support Systems in Irrigated Agriculture: A Review." *Agricultural Water Management* 257: 107161. doi: 10.1016/j.agwat.2021.107161.

Bale, Jeffery S., Gregory J. Masters, Ian D. Hodkinson, Caroline Awmack, T. Martijn Bezemer, Valerie K. Brown, Jennifer Butterfield, et al. 2002. "Herbivory in Global Climate Change Research: Direct Effects of Rising Temperature on Insect Herbivores." *Global Change Biology* 8 (1): 1–16. doi: 10.1046/j.1365-2486.2002.00451.x.

Barzman, Marco, Paolo Bàrberi, A. Nicholas E. Birch, Piet Boonekamp, Silke Dachbrodt-Saaydeh, Benno Graf, Bernd Hommel, et al. 2015. "Eight Principles of Integrated Pest Management." *Agronomy for Sustainable Development* 35 (4): 1199–1215. doi: 10.1007/s13593-015-0327-9.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. 2017. *Classification and Regression Trees*. New York: Routledge.

Bruneel, J., P. d'Este, and A. Salter. 2010. "Investigating the Factors That Diminish the Barriers to University–Industry Collaboration." *Research Policy* 39 (7): 858–868. doi: 10.1016/j.respol.2010.03.006.

Caffi, T., V. Rossi, and R. Bugiani. 2010. "Evaluation of a Warning System for Controlling Primary Infections of Grapevine Downy Mildew." *Plant Disease* 94 (6): 709–716. doi: 10.1094/PDIS-94-6-0709.

Caffi, T., S. E. Legler, V. Rossi, and R. Bugiani. 2012. "Evaluation of a Warning System for Early-Season Control of Grapevine Powdery Mildew." *Plant Disease* 96 (1): 104–110. doi: 10.1094/PDIS-06-11-0484.

Campello, R. J., D. Moulavi, and J. Sander. 2013. "Density-Based Clustering Based on Hierarchical Density Estimates." In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 160–172. Berlin, Heidelberg: Springer.

Cooreman, H., J. Vandenabeele, L. Debruyne, J. Ingram, H. Chiswell, A. Koutsouris, … F. Marchand. 2018. "A Conceptual Framework to Investigate the Role of Peer Learning Processes at on-Farm Demonstrations in the Light of Sustainable Agriculture." *International Journal of Agricultural Extension* 6 (3): 91–103.

Damos, P. 2015. “Modular Structure of Web-Based Decision Support Systems for Integrated Pest Management. A Review.” *Agronomy for Sustainable Development* 35 (4): 1347–1372. doi: 10.1007/s13593-015-0319-9.

Demirel, M., and N. A. Kumral. 2021. “Artificial Intelligence in Integrated Pest Management.” In *Artificial Intelligence and IoT-Based Technologies for Sustainable Farming and Smart Agriculture*, 289–313. Hershey, PA: IGI Global.

Durmuş, A., and A. Dağli. 2017. “Integration of Vocational Schools to Industry 4.0 by Updating Curriculum and Programs.” *International Journal of Multidisciplinary Studies and Innovative Technologies* 1 (1): 1–3.

Deutsch, C. A., J. J. Tewksbury, M. Tigchelaar, D. S. Battisti, S. C. Merrill, R. B. Huey, and R. L. Naylor. 2018. “Increase in Crop Losses to Insect Pests in a Warming Climate.” *Science (New York, N.Y.)* 361 (6405): 916–919. doi: 10.1126/science.aat3466.

Dara, S. K. 2019. “The New Integrated Pest Management Paradigm for the Modern Age.” *Journal of Integrated Pest Management* 10 (1): 12. doi: 10.1093/jipm/pmz010.

Directive 2009/128/EC of the European Parliament and of the Council of 21 October 2009 establishing a framework for Community action to achieve the sustainable use of pesticides. 2009, January 1. *Official Journal European Union Legislation L* 52 (309): 71–86.

Eitzinger, A., C. Binder, and M. Meyer. 2018. “Risk Perception and Decision-Making: Do Farmers Consider Risks from Climate Change?” *Climatic Change* 151 (3-4): 507–524. doi: 10.1007/s10584-018-2320-1.

European Commission 2020. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: A Farm to Fork Strategy for a Fair, Healthy and Environmentally-Friendly Food System.

Frank, E., M. A. Hall, and I. H. Witten. 2016. *The WEKA Workbench. Online Appendix for “Data Mining: Practical Machine Learning Tools and Techniques”*. 4th ed. San Francisco, California: Morgan Kaufmann.

Gent, D. H., E. De Wolf, and S. J. Pethybridge. 2011. “Perceptions of Risk, Risk Aversion, and Barriers to Adoption of Decision Support Systems and Integrated Pest Management: An Introduction.” *Phytopathology* 101 (6): 640–643. doi: 10.1094/PHYTO-04-10-0124.

Gent, D. H., W. F. Mahaffee, N. McRoberts, and W. F. Pfender. 2013. “The Use and Role of Predictive Systems in Disease Management.” *Annual Review of Phytopathology* 51: 267–289. doi: 10.1146/annurev-phyto-082712-102356.

Hernandez, J. E., J. Kacprzyk, H. Panetto, A. Fernandez, S. Liu, A. Ortiz, and M. De-Angelis. 2017. “Challenges and Solutions for Enhancing Agriculture Value Chain Decision-Making. A Short Review.” In *Working Conference on Virtual Enterprises*, 761–774. Cham: Springer.

Hochman, Z., and P. S. Carberry. 2011. “Emerging Consensus on Desirable Characteristics of Tools to Support Farmers’ Management of Climate Risk in Australia.” *Agricultural Systems* 104 (6): 441–450. doi: 10.1016/j.agsy.2011.03.001.

Hunter, J. D. 2007. “Matplotlib: A 2D Graphics Environment.” *Computing in Science & Engineering* 9 (3): 90–95. doi: 10.1109/MCSE.2007.55.

Ingram, J., H. Chiswella, J. Mills, L. Debruyne, H. Cooreman, A. Koutsouris, … F. Marchand. 2018. “Enabling Learning in Demonstration Farms: A Literature Review.” *International Journal of Agricultural Extension* 6 (3): 29–42.

Introne, J., and L. Iandoli. 2014. “Improving Decision-Making Performance through Argumentation: An Argument-Based Decision Support System to Compute with Evidence.” *Decision Support Systems* 64: 79–89. doi: 10.1016/j.dss.2014.04.005.

Johnen, A., and H. Meier. 2000. *A weather-based decision support system for managing oilseed rape pests.* Paper presented at the Brighton Crop Protection Conference Pests and Diseases.

Jones, V. P., J. F. Brunner, G. G. Grove, B. Petit, G. V. Tangren, and W. E. Jones. 2010. “A Web-Based Decision Support System to Enhance IPM Programs in Washington Tree Fruit.” *Pest Management Science* 66 (6): 587–595. doi: 10.1002/ps.1913.

Kabir, M. H., and R. Rainis. 2015. “Do Farmers Not Widely Adopt Environmentally Friendly Technologies? Lesson from Integrated Pest Management (IPM).” *Modern Applied Science* 9 (3): 208. doi: 10.5539/mas.v9n3p208.

Kanatas, P., I. S. Travlos, I. Gazoulis, A. Tataridas, A. Tsekoura, and N. Antonopoulos. 2020. “Benefits and Limitations of Decision Support Systems (DSS) with a Special Emphasis on Weeds.” *Agronomy* 10 (4): 548. doi: 10.3390/agronomy10040548.

Kluyver, T., B. Ragan-Kelley, F. Pérez, B. E. Granger, M. Bussonnier, J. Frederic, … C. Willing. 2016. *Jupyter Notebooks - a Publishing Format for Reproducible Computational Workflows*. Elpub, 2016: 87–90. doi: 10.3233/978-1-61499-649-1-87.

Kogan, M. 1998. “Integrated Pest Management: Historical Perspectives and Contemporary Developments.” *Annual Review of Entomology* 43: 243–270. doi: 10.1146/annurev.ento.43.1.243.

Kuehne, G., R. Nettle, and R. Llewellyn. 2019. “Key Social Processes Sustaining the Farmer/Advisor Relationship.” *Rural Extension and Innovation Systems Journal* 15 (1): 20–29.

Lacoste, M., and S. Powles. 2016. “Beyond Modelling: Considering User-Centred and Post-Development Aspects to Ensure the Success of a Decision Support System.” *Computers and Electronics in Agriculture* 121: 260–268. doi: 10.1016/j.compag.2016.01.003.

Lawrence, M., P. Goodwin, and R. Fildes. 2002. “Influence of User Participation on DSS Use and Decision Accuracy.” *Omega* 30 (5): 381–392. doi: 10.1016/S0305-0483(02)00048-8.

Lázaro, E., D. Makowski, and A. Vicent. 2021. “Decision Support Systems Halve Fungicide Use Compared to Calendar-Based Strategies without Increasing Disease Risk.” *Communications Earth & Environment* 2 (1): 224. doi: 10.1038/s43247-021-00291-8.

Lesecq, S., G. Keeley, G. Dudnik, and A. O’Riordan. 2022, June 30–July 2. “Would Digitalisation Support Transition towards More Sustainable Agriculture to Meet the Green Deal Ambition? Paper Presented at the.” 2022 8th International Conference on Energy Efficiency and Agricultural Engineering (EE&AE). doi: 10.1109/EEAE53789.2022.9831296.

Mahaman, B. D., H. C. Passam, A. B. Sideridis, and C. P. Yialouris. 2003. “DIARES-IPM: A Diagnostic Advisory Rule-Based Expert System for Integrated Pest Management in Solanaceous Crop Systems.” *Agricultural Systems* 76 (3): 1119–1135. doi: 10.1016/s0308-521x(02)00187-7.

Maraveas, C. 2022. "Incorporating Artificial Intelligence Technology in Smart Greenhouses: Current State of the Art." *Applied Sciences* 13 (1): 14. doi: 10.3390/app13010014.

Marić, M., I. Orović, and S. Stanković. (2016, May 30 - June 3). "Compressive Sensing Based Image Processing in TrapView Pest Monitoring System." Paper presented at the *2016 39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 508–512. 2016, Opatija, Croatia: IEEE.

McInnes, L., J. Healy, and J. Melville. 2018. "Umap: Uniform Manifold Approximation and Projection for Dimension Reduction." *arXiv Preprint arXiv:1802.03426 v3*. 1–63. doi: 10.48550/arXiv.1802.03426.

McKinney, W. 2010, June. "Data Structures for Statistical Computing in Python." In Proceedings of the 9th Python in Science Conference. Vol. 445, pp. 51–56 doi: 10.25080/Majora-92bf1922-00a.

Metzger, M. J., R. G. H. Bunce, R. H. Jongman, C. A. Mücher, and J. W. Watkins. 2005. "A Climatic Stratification of the Environment of Europe." *Global Ecology and Biogeography* 14 (6): 549–563. doi: 10.1111/j.1466-822X.2005.00190.

Moulavi, D., P. A. Jaskowiak, R. J. Campello, A. Zimek, and J. Sander. 2014, April. "Density-Based Clustering Validation." In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pp. 839–847. Society for Industrial and Applied Mathematics. doi: 10.1137/1.9781611973440.96.

Newman, S., T. Lynch, and A. A. Plummer. 2000. "Success and Failure of Decision Support Systems: Learning as we Go." *Journal of Animal Science* 77 (E-Suppl): 1–12. doi: 10.2527/jas2000.77E-Suppl1e.

Norton, G. A., D. Adamson, L. G. Aitken, L. J. Bilston, J. Foster, B. Frank, and J. K. Harper. 1999. "Facilitating IPM: The Role of Participatory Workshops." *International Journal of Pest Management* 45 (2): 85–90. doi: 10.1080/096708799227860.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, … E. Duchesnay. 2011. "Scikit-Learn: Machine Learning in Python." *The Journal of Machine Learning Research* 12: 2825–2830.

Quinlan, J. R. 1993. "Program for Machine Learning." *C4. 5*

Rose, David C., William J. Sutherland, Caroline Parker, Matt Lobley, Michael Winter, Carol Morris, Susan Twining, Charles Ffoulkes, Tatsuya Amano, and Lynn V. Dicks. 2016. "Decision Support Tools for Agriculture: Towards Effective Design and Delivery." *Agricultural Systems* 149: 165–174. doi: 10.1016/j.agsy.2016.09.009.

Rose, D. C., C. Parker, J. Fodery, C. Park, W. J. Sutherland, and L. V. Dicks. 2018. "Involving Stakeholders in Agricultural Decision Support Systems: Improving User-Centred Design." *International Journal of Agricultural Management* 6 (1029-2019-924): 80–89.

Rossi, V., T. Caffi, and F. Salinari. 2012. "Helping Farmers Face the Increasing Complexity of Decision-Making for Crop Protection." *Phytopathologia Mediterranea*: 51 (3): 457–479.

Rossi, V., F. Salinari, S. Poni, T. Caffi, and T. Bettati. 2014. "Addressing the Implementation Problem in Agricultural Decision Support Systems: The Example of Vite. net®." *Computers and Electronics in Agriculture* 100: 88–99. doi: 10.1016/j.compag.2013.10.011.

Rossi, V., Sperandio, G., Caffi, T., Simonetto, A., & Gilioli, G. (2019). Critical Success Factors for the Adoption of Decision Tools in IPM. *Agronomy*, 9(11), 710. doi: 10.3390/agronomy9110710.

Sciarretta, Andrea, Maria Rosaria Tabilio, Armando Amore, Marco Colacci, Miguel Á. Miranda, David Nestel, Nikos T. Papadopoulos, and Pasquale Trematerra. 2019. "Defining and Evaluating a Decision Support System (DSS) for the Precise Pest Management of the Mediterranean Fruit Fly, *Ceratitis Capitata*, at the Farm Level." *Agronomy* 9 (10): 608. doi: 10.3390/agronomy9100608.

Shtienberg, D. 2013. "Will Decision-Support Systems Be Widely Used for the Management of Plant Diseases?" *Annual Review of Phytopathology* 51: 1–16. doi: 10.1146/annurev-phyto-082712-102244.

Shukla, P. R., J. Skeg, E. C. Buendia, V. Masson-Delmotte, H. O. Pörtner, D. C. Roberts, … J. Malley. 2019. "Climate Change and Land: An IPCC Special Report on Climate Change, Desertification, Land Degradation, Sustainable Land Management, Food Security, and Greenhouse Gas Fluxes in Terrestrial Ecosystems."

Skendžić, S., M. Zovko, I. P. Živković, V. Lešić, and D. Lemić. 2021. "The Impact of Climate Change on Agricultural Insect Pests." *Insects* 12 (5): 1–31. doi: 10.3390/insects12050440.

Steffek, R., H. Reisenzein, G. Strauss, T. Leichtfried, J. Hofrichter, I. Kopacka, … R. Alt. 2011. "VitisCLIM, a Project Modelling Epidemiology and Economic Impact of Grapevine 'Flavescence Doree' Phytoplasma in Austrian Viticulture under a Climate Change Scenario." *Bulletin of Insectology* 64 (1): S191–S192.

Stenberg, J. A. 2017. "A Conceptual Framework for Integrated Pest Management." *Trends in Plant Science* 22 (9): 759–769. doi: 10.1016/j.tplants.2017.06.010.

Sudharsan, D., J. Adinarayana, A. K. Tripathy, S. Ninomiya, M. Hirafuji, T. Kiura, U. B. Desai, S. N. Merchant, D. Raji Reddy, and G. Sreenivas. 2012. GEOSENSE: A Multimode Information and Communication System. *ISRN Sensor Networks* 2012: 1–13. doi: 10.5402/2012/215103.

Taechatanasat, P., and L. Armstrong. 2014. "Decision Support System Data for Farmer Decision Making."

Tamayo, R. A. C., M. L. Ibarra, and J. A. G. Macías. 2010, September. "Better Crop Management with Decision Support Systems Based on Wireless Sensor Networks." In *2010 7th International Conference on Electrical Engineering Computing Science and Automatic Control*, 412–417. IEEE.

Taylor, R. A. J., D. A. Herms, J. Cardina, and R. H. Moore. 2018. "Climate Change and Pest Management: Unanticipated Consequences of Trophic Dislocation." *Agronomy* 8 (1): 7. doi: 10.3390/agronomy8010007.

Voogt, J., M. Almekinders, J. van den Akker, and B. Moonen. 2005. "A 'Blended'in-Service Arrangement for Classroom Technology Integration: Impacts on Teachers and Students." *Computers in Human Behavior* 21 (3): 523–539. doi: 10.1016/j.chb.2004.10.003.

Waskom, M. L. 2021. "Seaborn: Statistical Data Visualization." *Journal of Open Source Software* 6 (60): 3021. doi: 10.21105/joss.03021.

Witten, I. H., E. Frank, M. A. Hall, and C. J. Pal. 2016. "Data Mining Fourth Edition: Practical Machine Learning Tools and Techniques."

Zhou, H., T. T. Yuen, C. Popescu, A. Guillen, and D. G. Davis. 2015, April. "Designing Teacher Professional Development Workshops for Robotics Integration across Elementary and Secondary School Curriculum." In 2015 International Conference on Learning and Teaching in Computing and Engineering (215–216. IEEE.

Zou, K. H., J. R. Fielding, S. G. Silverman, and C. M. Tempany. 2003. "Hypothesis Testing I: Proportions." *Radiology* 226 (3): 609–613. doi: 10.1148/radiol.2263011500.

Appendix A: Farmers

The attributes "Country", "Region" and "Cluster" were not used for dimensionality reduction or clustering. The attribute "Region" was created from the "Country" data and used to divide the dataset into three separate datasets (i.e. farmers from Northern Europe, Central Europe and Southern Europe). The attribute "Cluster" was added after clustering and used as a target variable for the classification trees.

Table A1. Attributes, questions to which they refer and possible answers in the questionnaire for farmers.

		Possible answers (values)				
Attribute	Question related to the attribute	1	2	3	4	5
Age	Which is your age group?	18 to 25	25 to 35	35 to 50	50 to 60	60 or more
Already used DSS*	Have you already used decision support systems?	Yes	No			
Appreciate change*	On a scale from 1 to 7, how much do you appreciate change in general?	1, + 2, + 3, + 4 = 1 (not important)	5, + 6 = 2 (important to some extent)	7 = 3 (very important)		
Clusters	►Result of clustering (HDBSCAN). Attribute was added after dimensionality reduction and clustering.	−1 (outlier)	0	1		
Computer at home	Do you have a computer at home?	Yes	No			
Country	In which country are you based?	►Open question. The survey was conducted in 12 European countries: Denmark, Finland, France, Germany, Greece, Italy, Lithuania, Netherlands, Slovenia, Sweden, United Kingdom, and Norway. There were no farmers present on workshop in Norway.				
Degree	What is the highest degree you have attained?	Others	Vocational	Bachelor	Masters	PhD
DSS must be tailored*	Would you consider it necessary that a decision support system is tailored to your farm's characteristics?	Not necessary	Yes, if possible	Required		
Enjoy using new technology*	On a scale from 1 to 7, how much do you enjoy using new technologies?	1, + 2, + 3, + 4 = 1 (not important)	5, + 6 = 2 (important to some extent)	7 = 3 (very important)		
Farm size*	Your farm is:	Under 5 ha	5-10 ha	10-30 ha	30-60 ha	over 60 ha
Followed manufacturer's demo*	Have you followed a manufacturer's demonstration of a decision support system?	Never	Occasionally by invitation	Regularly		
Gender*	What is your gender?	Male	Female	Non-binary	I prefer not to say	
Importance of low price*	Please rate from 1 (not important at all) to 7 (very important) the importance of "low price" for a decision support system.	1, + 2, + 3, + 4 = 1 (not important)	5, + 6 = 2 (important to some extent)	7 = 3 (Very important)		
Importance of easy to use*	Please rate from 1 (not important at all) to 7 (very important) the importance of "easy to use" for a decision support system.	1, + 2, + 3, + 4 = 1 (not important)	5, + 6 = 2 (important to some extent)	7 = 3 (very important)		
Importance of efficiency*	Please rate from 1 (not important at all) to 7 (very important) the importance of "efficiency" for a decision support system.	1, + 2, + 3, + 4 = 1 (not important)	5, + 6 = 2 (important to some extent)	7 = 3 (very important)		
Importance of polyvalence*[1]	Please rate from 1 (not important at all) to 7 (very important) the importance of "polyvalence" for a decision support system.	1, + 2, + 3, + 4 = 1 (not important)	5, + 6 = 2 (important to some extent)	7 = 3 (very important)		

(continued)

Table A1. Continued.

		Possible answers (values)				
Attribute	Question related to the attribute	1	2	3	4	5
Income*	What is your average yearly income from your farming activity?	Less than 20.000 €	Between 20.000 € and 35.000 €	Between 35.000 € and 50.000 €	Between 50.000 € and 75.000 €	More than 75.000 €
Legislative requirements*	Do you need to satisfy some legislative or market requirements (e.g. limits of fertilizers use, time of pesticide applications, …)?	Yes	No			
Need IT teaching*	Would you need additional computer-related education?	No	If available, yes	I need it urgently		
Price important factor	Is price an important factor of your decision to use or not a given decision support system?	Yes	No			
Productivity improvement*	Do you think that the use of a decision support tool may improve your decision-making process and productivity?	No, not at all.	I am not sure	Probably	Yes, for sure	
Region	►Countries from attribute "Country" were combined in to three regions.	Northern Europe	Central Europe	Southern Europe		
High-speed internet*	Do you have access to high-speed internet?	No	Yes, in the office but not in the fields	Yes, both in the office and in the fields		
Trust in DSS*	On a scale from 1 to 7, how much would you trust the guidance of a decision support system?	1, + 2, + 3, + 4 = 1 (not important)	5, + 6 = 2 (important to some extent)	7 = 3 (very important)		
Willing to try new products*	Are you willing to try new products and services?	Yes	Sometimes	Never		
WTP for DSS*	Are you willing to pay in order to use a decision support system?	No	Only if a subsidy is available	Yes		

*Statistically significant differences between farmers classified in at least two clusters.

+ indicates separate answers in the questionnaire that were combined for the purposes of this research.

► marks important note and does not describe question or answer(s).

Appendix B: Farm advisors

The attributes "Country", "Region" and "Cluster" were not used for dimensionality reduction or clustering. The attribute "Region" was created from the "Country" data and used to divide the dataset into three separate datasets (i.e. farm advisors from Northern Europe, Central Europe and Southern Europe). The attribute "Cluster" was added after clustering and used as a target variable for the classification trees.

Table B1. Attributes, questions to which they refer and possible answers in the questionnaire for farm advisors.

		Possible answers (values)			
Attribute	Question related to the attribute	1	2	3	4
Access to marketing info about DSS[2]*	Do you have access to marketing information about decision support systems?	No	Occasionally only	Yes, I'm subscribed to receive news concerning decision support systems	
Age*	Which is your age group?	18 to 25, + 25 to 35	35 to 50, + 50 to 60	60 or more	
Clusters	►Result of clustering (HDBSCAN). Attribute was added after dimensionality reduction and clustering.	0	1	2	
Country	In which country are you based?	►Open question. The survey was conducted in 12 European countries: Denmark, Finland, France, Germany, Greece, Italy, Lithuania, Netherlands, Slovenia, Sweden, United Kingdom, and Norway.			
Crop type group*	What is the prevailing crop type of your clients?	Arable crops + Vegetables + Minor crops + Flowers	Vineyards + Orchards	Other + mixed categories	
Degree	What is the highest degree you have attained?	Others (please specify), + Vocational qualification	Bachelor's degree, + Master's degree, + PhD		
Evidence required	Do you require evidence-based proofs from developers?	No	Sometimes	If available always	
Experience*	How long have you been working in farm consulting?	Less than 2 years, + Between 2 and 5 years	Between 5 and 10 years, + More than 10 years		
Farm scale	What is the farm scale structure of the majority of your clients?	Small family farms (0 – 10 ha)	Midsize farms (10 - 30 ha)	Large farms (> 30 ha)	Mixed
Followed manufacturer's demo*	Have you followed a manufacturer's demonstration of a decision support system?	Never	Occasionally by invitation	Regularly	
Gender	What is your gender?	Male	Female	Non-binary	I prefer not to say
How often do you use DSS	How often do you use decision support systems to find the best pest management strategy in a given situation?	Never	Rarely + Sometimes	often + very often	
Legislative requirements	Do your clients need to satisfy some legislative or market requirements (e.g. limits of fertilizers use, time of pesticide applications, …)?	Yes	No		
Output preference*	Is a simplified graphical output from a decision support system enough for your needs?	Yes	No, both graphic and textual outputs are needed		
Productivity improvement*	Does the use of a decision support system improve your decision-making process and productivity?	No use of DSS at all, + No improvement	Some improvement	Great improvement	
Region	►Countries from attribute "Country" were combined into three regions.	Northern Europe	Central Europe	Southern Europe	
High-speed internet*	Do you have access to high-speed internet?	No	In the office but not in all the fields of my clients	Yes, both in the office and in the fields of my clients	
Substitute or complement	Do you think decision support systems are a substitute or complement to your job?	Complement	Substitute		

*Statistically significant differences between farm advisors classified in at least two clusters.
\+ indicates separate answers in the questionnaire that were combined for the purposes of this research.
► marks important note and not question or answer(s).

Chapter 4

Integrating Computer Vision in Marine Biology

This chapter presents the contributions related to the development of an accessible computer vision method with high confidence for phytoplankton biology.

Phytoplankton plays a key role in the production of organic matter and the carbon cycle in the sea. Its diversity influences the structure and processes of marine ecosystems, with the size and shape of the cells being crucial for nutrient uptake. Climate change, which manifests itself in increased seawater temperatures and changing river patterns, leads to smaller phytoplankton cells, which can alter marine ecosystems and their functions.

To investigate how climatic shifts affect the physiology and community structure of phytoplankton, this work developed a specialised computer vision system that can accelerate and automate the monitoring and analysis of microscopic images of selected phytoplankton species. The computer vision system performs object detection and instance segmentation tasks. In object detection, objects in an image are classified and localised by drawing a bounding box around the objects, which facilitates the counting and identification of objects in an image. Instance segmentation involves detecting and delimiting the individual objects in an image, which enables precise measurement of the individual objects. For marine biologists, visual explanation and accessibility through transfer learning improves the confidence of the ML system.

The approach uses transfer learning, in which generic, pre-trained YOLOv8 computer vision models are fine-tuned with microscope images from the Adriatic Sea. The development was done in Python within JupyterLab, the latest interface of the Jupyter project (Kluyver et al., 2016), the YOLOv8 computer vision library, the scikit-learn library for ML (Pedregosa et al., 2012), pandas (McKinney et al., 2010) and for visualisation seaborn (Waskom, 2021) and Matplotlib (Hunter, 2007).

The results of this study demonstrate the effectiveness of using computer vision techniques to identify phytoplankton and estimate biovolume. After fine-tuning, the YOLOv8 Nano models showed high accuracy in the detection and segmentation of phytoplankton cells from the microscopic images of the two species. High mAPs indicate that the model could accurately localise and classify phytoplankton species and that the segmentation masks generated by the model were precise and provided an accurate estimate of cell surface area, which is required for phytoplankton biovolume estimation.

This study demonstrates the potential of computer vision techniques, in particular instance segmentation and object detection, in analysing phytoplankton community structure. The study shows that it is possible to fine-tune a generic pre-trained model trained on out-of-domain images, and that this is possible even with a relatively small training dataset. Compared to manual analysis, the developed method significantly increases the number of samples that can be processed. This increase in sample size allows for a more robust and ecologically relevant assessment of phytoplankton community structure. The approach reduces the dependence on experienced taxonomists and makes it a scalable solution for long-term ecological monitoring.

This work directly addresses the following hypotheses defined in the introduction:

H5: Effective computer vision models for marine biology research can be developed via transfer learning on limited training data, and their trustworthiness increased with explainability methods.

Hypotheses H5 is fully confirmed by the results presented in this chapter.

The complete work is presented in the following publication and is included below:

Marzidovšek, M., Mozetič, P., Francé, J., & Podpečan, V. (2024). Computer Vision Techniques for Morphological Analysis and Identification of Two Pseudo-nitzschia Species. *Water*, 16(15), 2160.

The work was also presented at the 6th Biennial Conference of the Nordic Society Oikos (NSO) in Lund, Sweden, in 2024 and included in the publication:

Marzidovšek, M., Mozetič, P., Francé, J., & Podpečan, V. (2024). Analyzing the phytoplankton community structure with computer vision. *Nordic Oikos 2024: Abstract Book*, 83.

Martin Marzidovšek contributed to this work as follows: conceptualisation, methodology, software, formal analysis, writing —original draft, writing — review and editing, visualisation.

water

MDPI

Article

Computer Vision Techniques for Morphological Analysis and Identification of Two *Pseudo-nitzschia* Species

Martin Marzidovšek [1,2,*,†], Patricija Mozetič [3], Janja Francé [3] and Vid Podpečan [1]

[1] Jožef Stefan Institute, Knowledge Technologies E8, 1000 Ljubljana, Slovenia; vid.podpecan@ijs.si
[2] Jožef Stefan International Postgraduate School, 1000 Ljubljana, Slovenia
[3] National Institute of Biology, Marine Biology Station Piran, 6330 Piran, Slovenia; patricija.mozetic@nib.si (P.M.); janja.france@nib.si (J.F.)
* Correspondence: martin.marzidovsek@ijs.si
† Current address: Department of Knowledge Technologies, Jamova cesta 39, 1000 Ljubljana, Slovenia.

Abstract: The diversity of phytoplankton influences the structure of and processes that occur in marine ecosystems, with size and other morphological traits being crucial for nutrient uptake and retention in the euphotic zone. Our research introduces a machine learning method that can facilitate the analysis of phytoplankton functional traits from image data. We use computer vision to identify and quantify phytoplankton species and estimate size-related traits based on cell morphology. The study uses transfer learning, where generic, pre-trained YOLOv8 computer vision models are fine-tuned with microscope image data from the Adriatic Sea. The study shows that, for this task, it is possible to effectively fine-tune models trained on out-of-domain images and that this is possible with a small training dataset. The results show high accuracy in detecting and segmenting phytoplankton cells from the microscopic images of the two selected phytoplankton taxa. For detection, the model achieves AP scores of 88.1% for *Pseudo-nitzschia cf. delicatissima* and 90.9% for *Pseudo-nitzschia cf. calliantha*, while for segmentation, the scores are 88.4% for *Pseudo-nitzschia cf. delicatissima* and 91.2% for *Pseudo-nitzschia cf. calliantha*. Compared to manual image analysis, the developed automatic method significantly increases the number of samples that can be processed.

Keywords: computer vision; transfer learning; machine learning; phytoplankton; cell size; marine biology

check for updates

Citation: Marzidovšek, M.; Mozetič, P.; Francé, J.; Podpečan, V. Computer Vision Techniques for Morphological Analysis and Identification of Two *Pseudo-nitzschia* Species. *Water* **2024**, *16*, 2160. https://doi.org/10.3390/w16152160

Academic Editors: Guangyi Wang and José Luis Sánchez-Lizaso

Received: 14 June 2024
Revised: 25 July 2024
Accepted: 30 July 2024
Published: 31 July 2024

1. Introduction

Phytoplankton are marine microorganisms that are crucial for the production of organic matter and the carbon cycle in the oceans and account for about 50% of global primary production [1]. These microorganisms play an important role in structuring pelagic food webs and the ocean's biological carbon pump [2] and their diversity and community structure are important indicators of the health of marine ecosystems. The diversity of phytoplankton influences the structure of and processes that occur within marine ecosystems, with cell size and size-related traits being critical factors for nutrient uptake and retention within the euphotic zone. Climate change, which is characterised by increased seawater temperatures and stratification, is leading to a trend towards smaller phytoplankton cells [3].

It is important to understand how changing physical and chemical conditions in the sea affect phytoplankton physiology and community structure, and how these changes disrupt marine ecosystems and their functions. Understanding these impacts is critical for assessing the health of aquatic ecosystems and predicting future changes in primary production, biomass and carbon export to the seafloor. For this reason, information on phytoplankton functional traits is crucial for aquatic ecologists to assess trait variability and link diversity to ecosystem functioning [4]. The body or cell size of microorganisms is considered a master trait, not only because other morphological traits can be derived from it, but also because it influences other general traits such as resource acquisition and defence [5].

Conventional microscopy techniques for determining the abundance, diversity and cell size of phytoplankton are time consuming and require experienced taxonomists. These limitations hinder the processing of the large amount of samples required for a comprehensive analysis.

Recently, the use of machine learning (ML) in aquatic ecology and marine biology has evolved significantly due to the need to process large-scale ecological data efficiently and accurately. Automated computer vision systems can now process large datasets of microscope images quickly and with high precision, reducing the reliance on human experts and making them invaluable tools for ecological monitoring. These advances have been made possible by deep neural networks (DNNs), which are able to learn complex patterns and representations from large datasets [6]. Deep learning, which encompasses a range of deep neural network architectures, has been critical to advances in computer vision. DNN architectures such as convolutional neural networks (CNNs) [7,8] have demonstrated exceptional capabilities in image analysis, enabling the automatic identification, classification and segmentation of objects. Transfer learning, a technique in which a model trained on a large dataset is fine-tuned on a smaller, domain-specific dataset (see [9,10] for reviews), has further increased the utility of DNNs in specialised domains where training data are scarce.

These technologies have been used for various tasks such as species identification, behavioural analysis and monitoring populations of marine organisms [11]. These applications have significantly expanded the scope and scale of marine research, allowing scientists to gain insights that were previously difficult to obtain manually. An earlier example of an ML approach comes from [12], who developed an automated classification system for phytoplankton using imaging-in-flow cytometry that combined various image features and a support vector machine and achieved a high classification accuracy. Recent studies have shown the effectiveness of deep learning models such as CNNs in analysing phytoplankton images [13]. These models can be trained on large datasets of labelled images to detect different phytoplankton species with high accuracy.

Ciranni et al. [14] provide a comprehensive analysis of the wide range of computer vision techniques and methods that have emerged to facilitate the automated analysis of small- to large-scale datasets of plankton images. One of the studies presented is by the authors of [15], who applied a deep learning-based object detection system for specimen classification on their PMID2019 dataset (10,819 microscopic phytoplankton images with 24 different categories) and compared baseline performances with state-of-the-art object detection models such as Faster R-CNNs, an FPN, a Single-Shot-Detector (SSD) [16], a YOLOv3 [17] and a RetinaNet [18]. When training and testing on the six most common classes and with IoU thresholds of 0.5 and 0.75, they achieved an average precision (AP) score of 90% regardless of the model, with the best performance achieved with YOLOv3 at 93.10%. When evaluated at an IoU of 0.75, YOLOv3 was the worst model with a score of 82.81%, while FPN was the best model with 89.19%. In another study, Luo et al. [13] used CNNs to analyse plankton images (23.4 million images with 108 classes divided into 38 groups) by performing segmentation and classification, for which they achieved an average precision of 84% and recall of 40% for all groups.

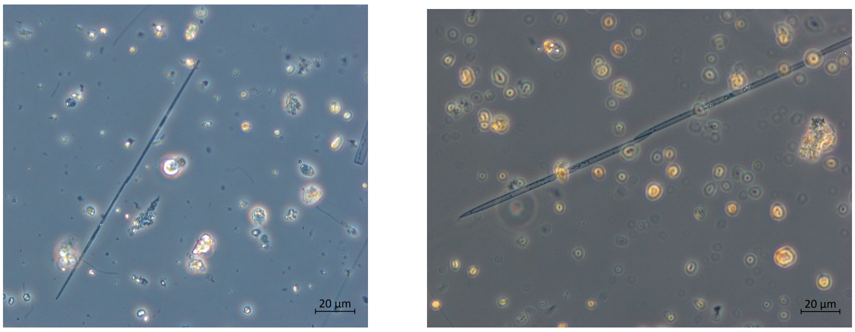
Despite the successes, there are still some challenges in the application of ML and computer vision in aquatic ecology. One of the biggest challenges is the need for large, annotated datasets to effectively train deep learning models. Annotating images is a time-consuming process that requires expert knowledge. Transfer learning can be a solution to this challenge as it can significantly reduce the training data requirements. Lysenko et al. [19] focused on the classification of Baikal phytoplankton using CNNs and achieved significant accuracy by using transfer learning.

As emphasised by [20], the ability of today's imaging systems can go well beyond only examining abundances of different taxa. According to them, the extraction of functional traits will be the next development in ML-based analysis of ecological images. They therefore suggest promising avenues for ML and computer vision approaches to extract information about functional traits from the images, such as biovolume estimation, which is of interest to our study.

The aim of this work is to develop a dedicated computer vision method that can accelerate and automate the monitoring and analysis of microscopic images of phytoplankton taxa. Besides the identification and localisation of phytoplankton species in image data, this method can also provide information on important traits such as cell size by predicting the cell area via instance segmentation with ML. The selection of the model objects—two species of *Pseudo-nitzschia*—was based on their ecological importance. These species, along with others from the same genus, are an important component of the phytoplankton community in the Gulf of Trieste [21,22]. Additionally, some species of this genus are potentially toxic, including the two selected taxa [23,24]. Therefore, improving methods for identifying species and distinguishing between potentially toxic and non-toxic species can significantly contribute to controlling seawater quality in designated aquaculture sites.

2. Materials and Methods

To train the computer vision model, a training dataset with labelled images had to be created. The microscopic images for this study were obtained from seawater samples collected in May/June 2023 from the southeastern part of the Gulf of Trieste (Adriatic Sea). A total of 306 images of phytoplankton were taken at 400× magnification using the Zeiss AxioObserver Z1 microscope with an integrated AxioCam Mrc5 digital camera. These images contained two species of interest, *Pseudo-nitzschia cf. delicatissima* (167 images) and *Pseudo-nitzschia cf. calliantha* (139 images), with the identification constrained by the limitations of light microscopy (Figure 1). For annotation, the distinction between the two taxa was based on the width of the transapical axis (smaller in *P. cf. delicatissima*) and on the characteristic overlapping of the cell ends in *P. cf. delicatissima*.



(**a**) (**b**)

Figure 1. Microscopic images of cells (elongated dark shaded objects) from the two phytoplankton species of interest *P. cf. delicatissima* (**a**) and *P. cf. calliantha* (**b**).

A total of 265 of the 306 microscopic images were manually annotated by identifying individual cell species, and drawing bounding boxes and segmentation masks. The open-source tool Computer Vision Annotation Tool (CVAT) [25] was used for this. These steps were crucial for the creation of a high-quality dataset for training the computer vision models. For cell area estimation, the model must predict a mask that covers the cell's surface, which requires accurate annotation masks. To evaluate its performance, the labelled dataset was split into a training (80%, 211 images) and a test dataset (20%, 54 images) using stratified random sampling to maintain the class distribution in both sets. In addition, the model internally performed data augmentation that included the following operators: hue, saturation, brightness, translation, scaling, flipping, and mosaic. Due to the small amount of labelled data, no tuning of the hyperparameters was performed.

To achieve the goal of the study, the computer vision system had to perform object detection and instance segmentation tasks. In object detection, objects in an image are classified and localised by drawing a bounding box around the objects, which facilitates the counting and identification of objects in an image. Instance segmentation involves detecting and delimiting the individual objects in an image, which enables precise measurement of the individual objects.

Due to the limited amount of training data (only 211 images for the two phytoplankton species), the approach involved fine-tuning two generic pre-trained models: one for object detection and one for instance segmentation. The models from the You Only Look Once (YOLO) series, in particular YOLOv8 (https://yolov8.com/, accessed on 29 July 2024), are among the most effective for these tasks. YOLO models are known for their speed and accuracy, making them suitable for real-time applications [17]. The architecture of the YOLOv8 model consists of (1) Backbone (CSPDarknet53, a CNN responsible for feature extraction), (2) Neck (C2f module that combines high-level semantic features with low-level spatial information) and (3) Head (making predictions) [26].

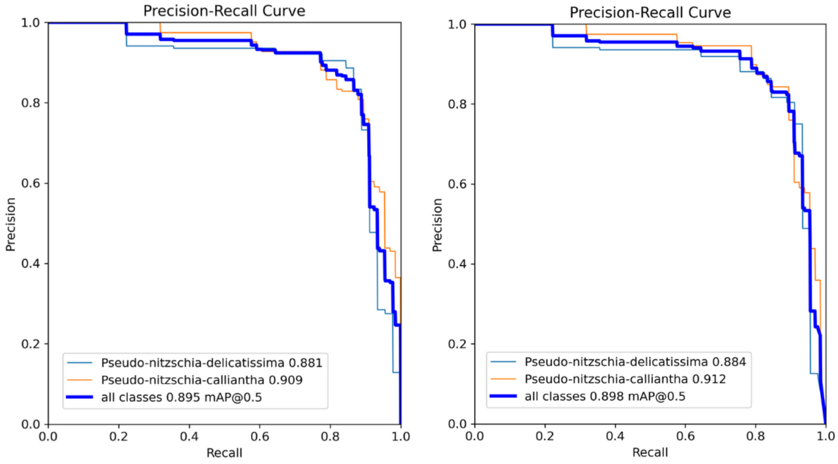
For this study, we selected the smallest Nano YOLOv8 models with 3.2 million (detection) and 3.4 million (segmentation) parameters. Both were pre-trained on the Common Objects in Context (COCO) dataset for large-scale object detection, segmentation and labelling [27]. Then, transfer-learning was used to fine-tune the pre-trained YOLOv8 model by re-training it on our annotated training dataset. This adapted the model to the specific characteristics of the selected phytoplankton species and improved its ability to accurately segment and detect individual phytoplankton cells. Key hyperparameters for training included a learning rate of 0.01, an image size of 1920 × 1920 pixels and a batch size of 16. The model was trained for 70 epochs, with training loss monitored to ensure convergence.

The performance of the model was evaluated using standard metrics for object detection and instance segmentation, including mean average precision (mAP) and intersection over union (IoU). IoU is a metric used to evaluate the accuracy of an object detector by comparing the overlap between the predicted bounding box and the ground truth bounding box. mAP is a crucial performance measure that captures the trade-off between precision at different levels of recall to evaluate the overall accuracy of object detection or segmentation models across multiple classes and selected IoU thresholds. These metrics provide a measure of the model's accuracy and precision in phytoplankton cell detection and segmentation. For an additional qualitative inspection of the results, a limited random sample (10% of the test set) of the predicted bounding boxes and segmentation masks was visualised and manually inspected to improve understanding of the models' results and trustworthiness with domain experts (see Figure 3).

The development was performed in Python within JupyterLab, the latest interface of the Jupyter project [28], YOLOv8 computer vision library, the scikit-learn machine learning library [29], pandas [30] and, for visualisations, seaborn [31] and Matplotlib [32].

3. Results

After fine-tuning, the YOLOv8 Nano models showed high accuracy in the detection and segmentation of phytoplankton cells from the microscopic images of the two species. For the detection, at an IoU of 0.5, the mAP for both species was 89.5%, and the AP was 88.1% for *Pseudo-nitzschia-delicatissima* and 90.9% for *Pseudo-nitzschia-calliantha*. For segmentation, the values were similar. The mAP for both species was 89.8%, and the AP was 88.4% for *Pseudo-nitzschia-delicatissima* and 91.2% for *Pseudo-nitzschia-calliantha*. At higher IoU thresholds (0.50 to 0.95), the mAP scores for both classes were 71.3% for detection and 51% for segmentation, respectively. Such mAPs indicate that the model could accurately localise and classify phytoplankton species and that the segmentation masks generated by the model were precise and provided an accurate estimate of cell area. Figure 2 shows the performance metrics for evaluating the model on the test set.



(**a**) detection (**b**) segmentation

Figure 2. Performance metrics of the model's evaluation on the test set for both tasks, (**a**) object detection and (**b**) instant segmentation. The area under the curve is the AP for each class and mAP for both classes.

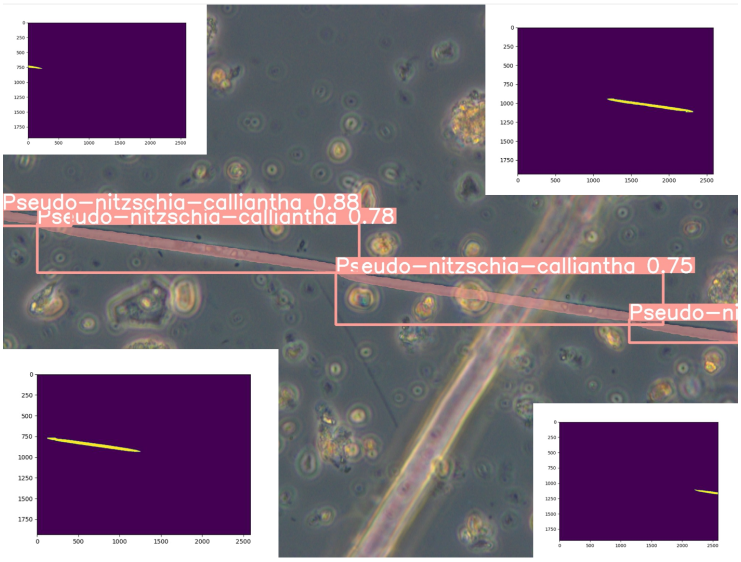


Figure 3. Visual inspection of the predicted classes, bounding boxes and segmentation masks. Separately drawn masks show the precision of the masks and how the model was able to discern difficult areas where cells chain together. While a few pixels in the difficult areas at the beginning and end of the cells are not correctly predicted, the most of the cells' bodies are drawn out correctly.

Furthermore, a visual inspection of the predicted bounding boxes and segmentation masks confirmed the precision of the models' results (Figure 3). It is important to note that the chaining of the cells of *P. cf. calliantha* and especially *P. cf. delicatissima* makes it difficult even for human annotators to distinguish one cell from another at the contact points where the cells chain together (Figure 3). This also makes the segmentation task very difficult for the computer vision model. Nevertheless, the model also performed quite well on this task, less optimal performance can be observed in the areas where the cells overlap since the model is struggling to discern to which cell instance a pixel belongs (Figure 3).

To investigate how the distributions of the actual cell area for both taxa correspond to the predicted ones, we superimposed the ground truth and predicted cell area distributions (Figure 4). The distributions of the ground truth (Figure 5) and the predicted cell area (Figure 4) show a similar pattern, indicating that the model is able to generalise well to new data. This indicates that the model can reliably estimate the cell area of phytoplankton.

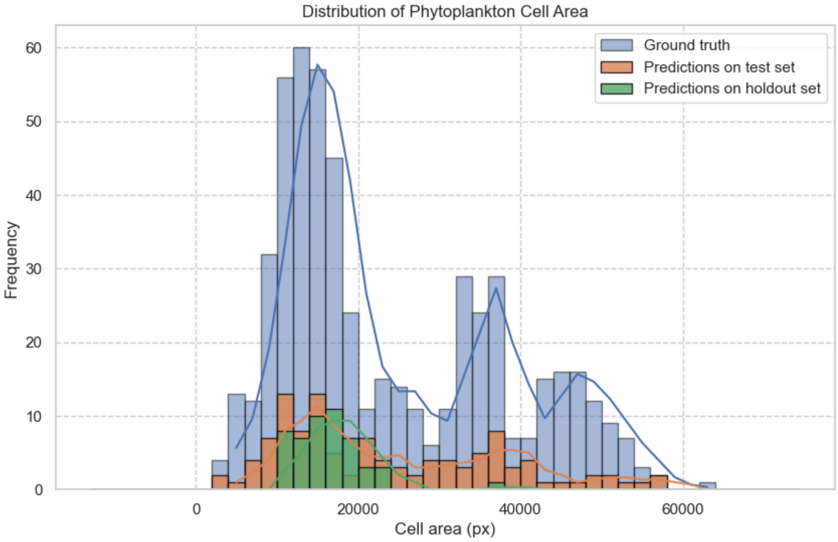


Figure 4. When superimposing distributions of the ground truth and predicted phytoplankton cell area it becomes apparent how the distributions of the areas of the predicted masks correspond to the actual cells area. This is a further indication of the quality of the model's predictions, as the distribution patterns are similar.

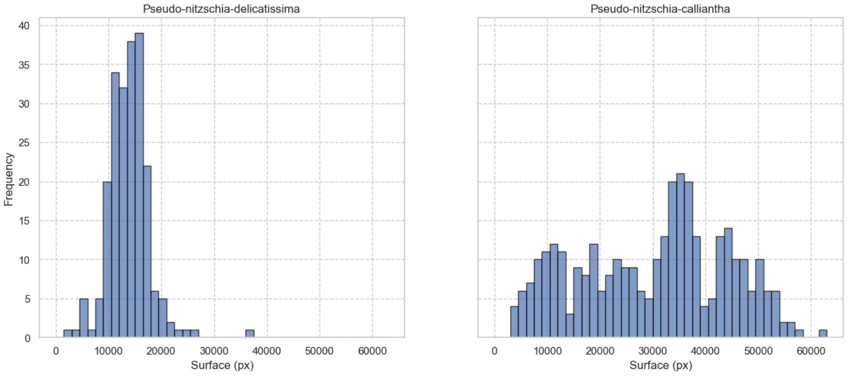


Figure 5. The two plots show how the actual cell areas are distributed for each of the *Pseudo-nitzschia* species. On the horizontal axis is the number of pixels representing the cell area and on the vertical axis is the number of cells with this cell area in the dataset. As you can see, *P. cf. delicatissima* has a more narrow and *P. cf. calliantha* a wider range of cell areas.

4. Discussion

The results of this study demonstrate the effectiveness of using computer vision techniques to identify phytoplankton and estimate important functional traits, such as cell size and biovolume. Our methodology closely matches the proposed protocols for characterising the functional traits of plankton from image data as described in a recent comprehensive review of ML techniques in [20]. The authors point out that biovolume and biomass can be estimated from the measured area of an organism or cell, and mention the application of segmentation algorithms when research requires pixel-level analysis. They also recommend evaluating the performance of the algorithms with the mAP score.

One of the challenges the study faced was the limited number of labelled images. This problem was addressed using transfer learning which allowed the model to leverage knowledge gained from pre-training on the COCO dataset. Our results show that fine-tuning of a generic pre-trained computer vision model is possible with a relatively small number of images, even if these images are from different domains (e.g., the COCO dataset does not contain microscopic images). This is in line with similar studies on plankton ecology where the benefits of using such approaches for classification have been reported (e.g., [33,34]).

Most similar published work is on marine plankton differentiation (classification) and detection; however, for comparisons to our study, the most suitable are those dealing with instance segmentation on plankton images, which are less abundant. The recent comprehensive review on automatic plankton image recognition by [35] only mentions the comparison in [36] of a semantic segmentation model SegNet [37], and an instance segmentation model Mask R-CNN [38] for algae detection and recognition.

Ruiz-Santaquiteria et al. [36] used Mask R-CNN, which is based on ResNet101 and was pre-trained on COCO for the instance segmentation of 126 plankton images (10 different taxa). Their best model for instance segmentation achieved an average precision of 85% with 86% recall. Unfortunately, the authors do not provide the mAP and IoU thresholds that would allow a more direct comparison. Ciranni et al. [14] summarise the work of [39] who compare three alternative approaches for plankton entity segmentation in flow-cytometry images. Similar to our approach, their Mask R-CNN model was pre-trained on COCO and they manually annotated over 3000 images. The method relying on the Mask R-CNN showed the best trade-off in terms of precision (16.3%) and recall (91.9%). In their study, Bergum et al. [40] also used a Mask R-CNN pre-trained on the COCO dataset and fine-tuned on their own dataset of 126 images of planktonic organisms (copepods). With their fine-tuned X101-32x8d-FPN-3x model, they reported an AP of 41.1% for instance segmentation.

Another study dealing with similar tasks of phytoplankton detection and instance segmentation in microscope image data from water samples was carried out by [41], who used classical ML algorithms and exploited texture and colour features in the images. Similarly to our approach, they used an IoU of 50% for a detection as a true positive. For phytoplankton detection, they reported that the best model was the random forest classifier using texture features with a precision of 77.2% and a recall of 90%.

In comparison, our model was fine-tuned with only 211 images and also included only two classes, which is less than most of the similar studies. Our model achieved an mAP score of 89.5% for detection and 89.8% for instance segmentation at an IoU of 0.5. At higher IoU thresholds (0.50 to 0.95), mAP scores were 71.3% for detection and 51% for segmentation, respectively. Although our main evaluation metric was mAP, we also report a precision of 84.9% and recall of 84.2% for instance segmentation to facilitate comparison with studies that reported these metrics, although they refer to classification evaluation, which is a simpler task than instance segmentation.

These comparisons underscore the effectiveness of our approach in processing instance segmentation tasks for extracting information about phytoplankton functional traits from microscopic images. The results show that while larger annotated datasets generally contribute to a higher accuracy, the combination of transfer learning and accurate annotation can yield competitive results even with a small number of annotated images. Furthermore, it was observed that, during training, the higher-resolution images improved the precision of segmentation tasks, which is likely applicable to real-world scenarios. The model's strong performance on these challenging tasks can also be attributed to the accurate mask annotations in the training set, though creating a large number of such annotations can be costly.

In our future work, we will focus on refining the model to solve difficult cases, such as overlapping cells, and improving the precision of segmentation masks, which is particularly relevant for cell area estimation. To this end, performance improvements should be sought at higher IoU settings during training, as this will ensure that more accurate segmentation masks are generated. Higher IoUs would place higher demands on the precision of the model in capturing the exact contours of thin and elongated objects such as the studied *Pseudo-nitzschia* species. However, this could also come at the cost of false negatives, meaning that some cells are missed altogether.

We also plan to expand the dataset by annotating more images, as this would likely further improve the performance of the model. We would also like to include other important phytoplankton species. Testing the model in real environments, will allow further validation and insight into its practical applications, where the approach can also be adapted for optimal cost–benefit.

5. Conclusions

Our study demonstrates the potential of computer vision techniques, particularly instance segmentation, in analysing phytoplankton functional traits, enabling aquatic ecologists to better understand their community structure. Our object detection model accurately detects and identifies the two phytoplankton from microscopic images. The instance segmentation model provides a good estimate of individual cell area, which is necessary for estimating another important functional trait, biovolume. The study shows that it is possible to fine-tune a generic pre-trained model trained on out-of-domain images, and that this is possible even with a relatively small training dataset.

Compared to manual analysis of phytoplankton images, the automated method significantly increases the number of samples to be processed. Such automation could provide a scalable solution to the spatially and temporally extensive monitoring programs.

Author Contributions: Conceptualization, M.M. and P.M.; methodology, M.M.; software, M.M.; validation, P.M. and J.F.; formal analysis, M.M.; investigation, P.M.; data curation, J.F.; writing—original draft preparation, M.M.; writing—review and editing, M.M., V.P., P.M. and J.F.; visualisation, M.M.; supervision, P.M. and V.P.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received funding from Slovenian Research and Innovation Agency (grant numbers P1-0237 and P2-0103).

Data Availability Statement: The complete code and data are available in an online repository: https://github.com/MartinMarzi/phytoplankton-cv4ecology, accessed on 29 July 2024.

Acknowledgments: We gratefully acknowledge the kind support of the California Institute of Technology and the Resnick Sustainability Institute, the mentors at CV4Ecology Summer Workshop Shir Bar, Tarun Sharma, Sara Beery, the Department of knowledge technologies—Jožef Stefan Institute (Sašo Džeroski), the Marine Biology Station—National Institute of Biology, and Tibor Frković for the annotations.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Falkowski, P.; Scholes, R.J.; Boyle, E.; Canadell, J.; Canfield, D.; Elser, J.; Gruber, N.; Hibbard, K.; Högberg, P.; Linder, S.; et al. The global carbon cycle: A test of our knowledge of earth as a system. *Science* **2000**, *290*, 291–296. [CrossRef] [PubMed]
2. Basu, S.; Mackey, K.R.M. Phytoplankton as Key Mediators of the Biological Carbon Pump: Their Responses to a Changing Climate. *Sustainability* **2018**, *10*, 869. [CrossRef]

3. Maranón, E.; Cermeno, P.; Latasa, M.; Tadonléké, R.D. Temperature, resources, and phytoplankton size structure in the ocean. *Limnol. Oceanogr.* **2012**, *57*, 1266–1278. [CrossRef]
4. Martini, S.; Larras, F.; Boyé, A.; Faure, E.; Aberle, N.; Archambault, P.; Bacouillard, L.; Beisner, B.E.; Bittner, L.; Castella, E.; et al. Functional trait-based approaches as a common framework for aquatic ecologists. *Limnol. Oceanogr.* **2021**, *66*, 965–994. [CrossRef]
5. Andersen, K.H.; Berge, T.; Gonçalves, R.J.; Hartvig, M.; Heuschele, J.; Hylander, S.; Jacobsen, N.S.; Lindemann, C.; Martens, E.A.; Neuheimer, A.B.; et al. Characteristic sizes of life in the oceans, from bacteria to whales. *Annu. Rev. Mar. Sci.* **2016**, *8*, 217–241. [CrossRef] [PubMed]
6. Lecun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 7553. [CrossRef] [PubMed]
7. LeCun, Y.; Boser, B.; Denker, J.S.; Henderson, D.; Howard, R.E.; Hubbard, W.; Jackel, L.D. Backpropagation applied to handwritten zip code recognition. *Neural Comput.* **1989**, *1*, 541–551. [CrossRef]
8. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
9. Pan, S.J.; Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **2009**, *22*, 1345–1359. [CrossRef]
10. Weiss, K.; Khoshgoftaar, T.M.; Wang, D. A survey of transfer learning. *J. Big Data* **2016**, *3*, 1–40. [CrossRef]
11. Irisson, J.O.; Ayata, S.D.; Lindsay, D.J.; Karp-Boss, L.; Stemmann, L. Machine learning for the study of plankton and marine snow from images. *Annu. Rev. Mar. Sci.* **2022**, *14*, 277–301. [CrossRef] [PubMed]
12. Sosik, H.M.; Olson, R.J. Automated taxonomic classification of phytoplankton sampled with imaging-in-flow cytometry. *Limnol. Oceanogr. Methods* **2007**, *5*, 204–216. [CrossRef]
13. Luo, J.Y.; Irisson, J.O.; Graham, B.; Guigand, C.; Sarafraz, A.; Mader, C.; Cowen, R.K. Automated plankton image analysis using convolutional neural networks. *Limnol. Oceanogr. Methods* **2018**, *16*, 814–827. [CrossRef]
14. Ciranni, M.; Murino, V.; Odone, F.; Pastore, V.P. Computer vision and deep learning meet plankton: Milestones and future directions. *Image Vis. Comput.* **2024**, *143*, 104934. [CrossRef]
15. Li, Q.; Sun, X.; Dong, J.; Song, S.q.; Zhang, T.; Liu, D.; Zhang, H.; Han, S. Developing a microscopic image dataset in support of intelligent phytoplankton detection using deep learning. *ICES J. Mar. Sci.* **2020**, *77*, 1427–1439. [CrossRef]
16. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016*; Springer International Publishing: Cham, Switzerland, 2016; pp. 21–37.
17. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
18. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988.
19. Lysenko, A.; Oznobikhin, M.; Kireev, E.; Dubrova, K.; Vorobyeva, S. Identification of Baikal phytoplankton inferred from computer vision methods and machine learning. In *Limnology and Freshwater Biology*; Limnological institute SB RAS: Irkutsk, Russia, 2021; pp. 1143–1146.
20. Orenstein, E.C.; Ayata, S.D.; Maps, F.; Becker, É.C.; Benedetti, F.; Biard, T.; de Garidel-Thoron, T.; Ellen, J.S.; Ferrario, F.; Giering, S.L.; et al. Machine learning techniques to characterize functional traits of plankton from image data. *Limnol. Oceanogr.* **2022**, *67*, 1647–1669. [CrossRef] [PubMed]
21. Cerino, F.; Fornasaro, D.; Kralj, M.; Giani, M.; Cabrini, M. Phytoplankton temporal dynamics in the coastal waters of the north-eastern Adriatic Sea (Mediterranean Sea) from 2010 to 2017. *Nat. Conserv.* **2019**, *34*, 343–372. [CrossRef]
22. Mozetič, P.; Cangini, M.; Francé, J.; Bastianini, M.; Aubry, F.B.; Bužančić, M.; Cabrini, M.; Cerino, F.; Čalić, M.; D'Adamo, R.; et al. Phytoplankton diversity in Adriatic ports: Lessons from the port baseline survey for the management of harmful algal species. *Mar. Pollut. Bull.* **2019**, *147*, 117–132. [CrossRef]
23. Arapov, J.; Ujević, I.; Straka, M.; Skejić, S.; Bužančić, M.; Bakrač, A.; Ninčević Gladan, Ž. First evidence of domoic acid production in Pseudo-nitzschia calliantha cultures from the central Adriatic Sea. *Acta Adriat.* **2020**, *61*, 135–144. [CrossRef]
24. Turk Dermastia, T.; Dall'Ara, S.; Dolenc, J.; Mozetič, P. Toxicity of the diatom genus Pseudo-nitzschia (Bacillariophyceae): Insights from toxicity tests and genetic screening in the northern Adriatic Sea. *Toxins* **2022**, *14*, 60. [CrossRef]
25. CVAT.ai Corporation. Computer Vision Annotation Tool (CVAT). 2024. Available online: https://www.cvat.ai/ (accessed on 29 July 2024).
26. Torres, J. YOLOv8 Architecture: A Deep Dive into Its Architecture. 2024. Available online: https://yolov8.org/yolov8-architecture/ (accessed on 11 June 2024).
27. Lin, T.; Maire, M.; Belongie, S.J.; Bourdev, L.D.; Girshick, R.B.; Hays, J.; Perona, P.; Ramanan, D.; Doll'ar, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. *arXiv* **2014**, arXiv:1405.0312.
28. Kluyver, T.; Ragan-Kelley, B.; Pérez, F.; Granger, B.E.; Bussonnier, M.; Frederic, J.; Kelley, K.; Hamrick, J.B.; Grout, J.; Corlay, S.; et al. Jupyter Notebooks-a publishing format for reproducible computational workflows. *Elpub* **2016**, *2016*, 87–90.
29. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Müller, A.; Nothman, J.; Louppe, G.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2012**, *12*, 2825–2830. arxiv:1201.0490.
30. McKinney, W. Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference, Austin, TX, USA, 28–30 June 2010; Volume 445, pp. 51–56.
31. Waskom, M. seaborn: Statistical data visualization. *J. Open Source Softw.* **2021**, *6*, 3021. [CrossRef]
32. Hunter. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **2007**, *9*, 90–95. [CrossRef]

33. Orenstein, E.C.; Beijbom, O. Transfer learning and deep feature extraction for planktonic image data sets. In Proceedings of the 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), Santa Rosa, CA, USA, 27–29 March 2017; pp. 1082–1088.
34. González, P.; Castaño, A.; Peacock, E.E.; Díez, J.; Del Coz, J.J.; Sosik, H.M. Automatic plankton quantification using deep features. *J. Plankton Res.* **2019**, *41*, 449–463. [CrossRef]
35. Eerola, T.; Batrakhanov, D.; Barazandeh, N.V.; Kraft, K.; Haraguchi, L.; Lensu, L.; Suikkanen, S.; Seppälä, J.; Tamminen, T.; Kälviäinen, H. Survey of automatic plankton image recognition: Challenges, existing solutions and future perspectives. *Artif. Intell. Rev.* **2024**, *57*, 114. [CrossRef]
36. Ruiz-Santaquiteria, J.; Bueno, G.; Deniz, O.; Vallez, N.; Cristobal, G. Semantic versus instance segmentation in microscopic algae detection. *Eng. Appl. Artif. Intell.* **2020**, *87*, 103271. [CrossRef]
37. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
38. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
39. Panaïotis, T.; Caray-Counil, L.; Woodward, B.; Schmid, M.S.; Daprano, D.; Tsai, S.T.; Sullivan, C.M.; Cowen, R.K.; Irisson, J.O. Content-aware segmentation of objects spanning a large size range: Application to plankton images. *Front. Mar. Sci.* **2022**, *9*, 870005. [CrossRef]
40. Bergum, S.; Saad, A.; Stahl, A. Automatic in-situ instance and semantic segmentation of planktonic organisms using Mask R-CNN. In Proceedings of the Global Oceans 2020: Singapore–US Gulf Coast, Biloxi, MS, USA, 5–30 October 2020; pp. 1–8.
41. Rivas-Villar, D.; Rouco, J.; Penedo, M.G.; Carballeira, R.; Novo, J. Automatic detection of freshwater phytoplankton specimens in conventional microscopy images. *Sensors* **2020**, *20*, 6704. [CrossRef] [PubMed]

Chapter 5

Network Analysis and Machine Learning on Graphs in Soil Ecology

This chapter presents the contributions that deal with the development of specialised software tools for network analysis and explainable ML on graph-structured ecological data.

In the field of ecology, a gap was identified for specialised tools for ecological research that provide network analysis metrics, interactive network visualisation and explainable ML on graphs. Therefore, BEFANA, a free and open source software tool, was developed to make common analysis tasks on ecological networks more accessible. The tool was developed in collaboration with ecology experts and is customised to the needs of ecologists. Its functionality enables them to analyse the topology and dynamics of ecological networks and apply selected ML algorithms. BEFANA offers methods for loading and pre-processing data, network analysis and interactive visualisation as well as modelling with experimental data. To utilise the network data for predictive and descriptive ML tasks, the data (nodes, edges) can be represented as numerical vectors or node embeddings using the node2vec ML algorithm (Grover & Leskovec, 2016). BEFANA enables the enrichment and modelling of networks with experimental data. The user can embed experimental or field data into the network structure, which can provide insight on the dynamics of the studied ecosystem.

BEFANA was evaluated by applying it to a soil food web of an agricultural grassland. This case study shows the main components and functions of BEFANA, and all the analysis steps are presented in the published work below.

BEFANA is based on an open source software stack which helps to achieve the following three goals: (a) ease of use, (b) interactivity and (c) extensibility. The software tools and ML pipelines are implemented in the Python programming language, which has become the de facto language of choice for scientific computing (Perez & Granger, 2007). In combination with JupyterLab, an interactive development environment for computational notebooks (i.e. Jupyter notebooks), and pandas (McKinney et al., 2010), it forms one of the best approaches for the dissemination of scientific programs, experiments, data manipulation and interactive computational notebooks. The ecological networks are encoded and analysed using NetworkX (Hagberg et al., 2008) and the node2vec algorithm (Grover & Leskovec, 2016) was used for representation learning. Seaborn (Waskom, 2021) and Matplotlib (Hunter, 2007) are used for visualisations.

Although it is common for computational notebooks to run locally, BEFANA goes one step further and also offers a public web interface via mybinder.org, a free public service that provides a community-driven infrastructure using BinderHub technology (Bussonnier et al., 2018), which allows the sharing of interactive code. Local installation of BEFANA is possible with Docker or by creating a virtual Python environment in which BEFANA can be launched.

The results of this work are as follows. BEFANA, a specialised software tool that enables ecologists to apply network analysis and ML algorithms to ecological networks. It helps them to quantify and test hypotheses about how the topology of a network influences its dynamics and how the functioning of the network in turn affects its structure.

In the case study, a hierarchical directed graph is constructed representing a detrital soil food web using one regional metaweb of all possible literature-based multitrophic links from a local pool of food webs (Mulder et al. 2005; Sechi et al., 2015). By applying network analysis, all nodes of the soil food web were ranked based on the out-degree to in-degree ratio (van den Brink and Rusinowska 2021), demonstrating that the web contains exactly one strongly- connected component spanning three nodes. Clustering separated the trophic groups based on the energy sources, which revealed that most trophic groups at the first level are vulnerable, in contrast to the trophic groups at the second and third trophic levels (average indegree is 8.2). Modelling with experimental data allowed a simple calculation of the total biomass for each of the trophic levels and resulted in a ranking of trophic groups by abundance. Earthworms and the three main resources for the plant-driven pathway, the bacteria-driven pathway and the fungus-driven pathway exhibit the highest biomass values. Representation learning shows that the topological information is indeed encoded in node embeddings, and dimensionality reduction followed by visualisation in 1D space gives a clearer indication of node similarity.

This work directly addresses the following hypothesis defined in the introduction:

H4: Specially developed software integrating accessible network analysis, ML, and explainability methods can enhance understanding and knowledge discovery from ecological networks.

Hypothesis H4 is fully confirmed by the results presented in this chapter.

The complete work is presented in the following publication and is included below:

Marzidovšek, M., Podpečan, V., Conti, E., Debeljak, M., & Mulder, C. (2022). BEFANA: A tool for biodiversity-ecosystem functioning assessment by network analysis. *Ecological Modelling*, 471, 110065.

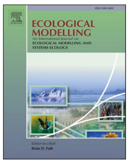
Martin Marzidovšek contributed to this work as follows: conceptualisation, methodology, software, formal analysis, writing the original draft, reviewing and editing, visualisation.

Ecological Modelling 471 (2022) 110065

Contents lists available at ScienceDirect

Ecological Modelling

journal homepage: www.elsevier.com/locate/ecolmodel



Short communication

BEFANA: A tool for biodiversity-ecosystem functioning assessment by network analysis

Martin Marzidovšek [a,b,*], Vid Podpečan [a], Erminia Conti [c], Marko Debeljak [a,b], Christian Mulder [c]

[a] *Jozef Stefan Institute, Department of Knowledge Technologies, University of Ljubljana, Slovenia*
[b] *Jožef Stefan International Postgraduate School, University of Ljubljana, Slovenia*
[c] *Department of Biological, Geological and Environmental Sciences, University of Catania, Italy*

ARTICLE INFO

Keywords:
Biodiversity
Ecology
Ecosystem
Food web
Graph theory
Soil ecology
Machine learning
Open-source software

ABSTRACT

BEFANA is a free and open-source software tool for ecological network analysis and visualisation. It is adapted to ecologists' needs and allows them to study the topology and dynamics of ecological networks as well as to apply selected machine learning algorithms. BEFANA is implemented in Python, and structured as an ordered collection of interactive computational notebooks. It relies on widely used open-source libraries, and aims to achieve simplicity, interactivity, and extensibility. BEFANA provides methods and implementations for data loading and pre-processing, network analysis and interactive visualisation, modelling with experimental data, and predictive modelling with machine learning. We showcase BEFANA through a concrete example of a detrital soil food web of one agricultural grassland, and demonstrate all of its main components and functionalities.

1. Introduction

Unsustainable use of natural resources have been compelling the scientific community to turn its interest to biodiversity loss in general, and to negative cascading effects in particular. Several works concerning ecological networks have been published[1] but too few of them are mechanistic and predictive, despite the plea by Ings et al. (2009). On the other hand, online dataset collections and international research interest is rapidly growing, especially for soil organisms (van den Hoogen et al. 2020; Phillips et al. 2021).

Data on the ecological interactions in food webs are generally represented in four different ways: I) Resource-to-consumer food web, where the structural model is either static or dynamic, with directed graphs (flow) for functional groups, emphasising energy channels, loops, biomass and nutrient cycling (Hunt and Wall 2002; Fitter 2005; Mulder et al. 2005); II) Allometric scaling (size spectra), where the resulting power law is often environmentally-driven, showing triangular plots for taxa, trophic generality and vulnerability, trophic heights, assimilation efficiency, addressing stoichiometric imbalances (Reuman and Cohen 2004; Mulder and Elser 2009); III) Topological networks, with a static biopyramid for species, trophospecies and/or functional groups, emphasising structural biodiversity, size-delimited pathways, nested hierarchy, and/or parasitism (Cohen et al., 2003; Cohen et al., 2009; Hudson et al. 2013; Mulder et al. 2005) and finally, IV) Interaction networks for species, trophospecies and/or functional groups, focusing on connectance, complexity, trophic transfer efficiency, self-damping, distribution of trophic links and chain lengths, magnitude of feeding relationships etc. (Reuman and Cohen 2004; Kondoh et al., 2010; Bascompte and Jordano 2007).

This work presents BEFANA, a free and open-source software tool suitable for ecological network analysis and visualisation for resource-to-consumer food webs, and topological and interaction networks (the aforementioned architectural types I, III and IV). It allows the ecologist to study the topology of the network and its temporal dynamics, identify indirect interactions and more. Ease of use, interactivity, and extensibility were the main guidelines when designing and developing BEFANA. As a result, BEFANA is implemented as a collection of interactive computational notebooks and relies on widely used open-source libraries for network analysis, data science and machine learning. BEFANA is available through a public web interface as well as a Docker

---

* Corresponding author at: Jozef Stefan Institute, Jamova 39, Ljubljana, 1000 Slovenia.
E-mail address: martin.marzidovsek@ijs.si (M. Marzidovšek).

[1] 443 papers recorded on Web of Science as of February 15th, 2022, using the specific query: AK=(ecological network*) AND AB=(ecological network*) AND TI=(ecological network*) NOT AK=(social).

https://doi.org/10.1016/j.ecolmodel.2022.110065
Received 5 May 2022; Received in revised form 29 June 2022; Accepted 3 July 2022

container and local installation.[2] The rest of the paper showcases selected parts of BEFANA through a case study of detrital soil food-web analysis.

2. Software design and implementation

2.1. Overview

The design of BEFANA (Fig. 1) follows the latest advances in developing and sharing of scientific software. It is based on the computational notebook paradigm and follows the "batteries included" philosophy by providing integration of libraries a user might need in ecological network analysis. As such, BEFANA is an interactive software tool but also a solid base for creating custom applications and experiments in this and related domains.

BEFANA is implemented in the Python programming language, which has become a *de facto* language of choice for scientific computing (Perez and Granger 2007). In combination with JupyterLab, an interactive development environment for computational notebooks (i.e. Jupyter notebooks), and pandas (McKinney, 2010), a Python library for data manipulation and analysis offering optimised data structures and functions, it forms one of the best approaches to the dissemination of scientific programs, experiments, data manipulation and interactive computational notebooks. BEFANA is based on an open-source software stack which helps to achieve the following three goals: (a) ease of use, (b) interactivity, and (c) extensibility.

The libraries currently included in BEFANA cover the following topics: data input and output, pre-processing and manipulation, charting, web-based graphical user interfaces, network analysis and visualisation, and machine learning. It is an open collection as any other libraries can be applied in BEFANA notebooks as long as they implement an interface to the Python interpreter or are available as a web service or a command line tool.

Even though it is common for computational notebooks to be run locally, BEFANA goes a step further and also provides a public web interface through *mybinder.org*, a free public service which provides a community-led infrastructure deploying the BinderHub technology (Project Jupyter et al. 2018), which allows sharing of interactive code. Local installation of BEFANA is possible using Docker or by creating a virtual Python environment where BEFANA can be started.

2.2. Software capabilities

BEFANA is a software tool for ecological network analysis and visualisation that allows to study the topology and dynamics of ecological networks as well as to apply selected machine learning algorithms. Its computational notebooks, code, interface and results, including diagrams and graphics, interweave and offer a uniquely-rich user experience. BEFANA can be used to demonstrate some particular implementation, to showcase an example, to share charts with results or to provide network visualisations. The software has been developed specifically for the purpose of soil food web analysis but can be easily adapted and extended for the analysis of other types of networks as well. The logical organisation of its functionalities is as follows.

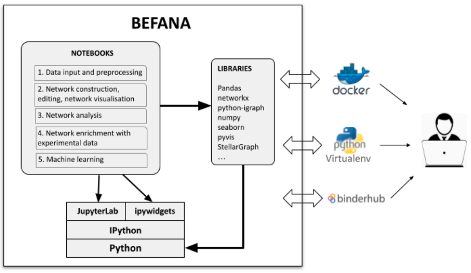


Fig. 1. Architecture of BEFANA.

2.2.1. Data loading and pre-processing

BEFANA supports several different input formats including csv, Excel, Access and HDF. The data is parsed and loaded into a high-performance data structure which supports advanced data selection and manipulation. If the input data requires pre-processing steps such as imputation, filtering, resampling, computing values etc. this is the right place to apply them.

2.2.2. Network construction

The data from which a network is constructed can be provided in several different ways. For example: (a) the network can be loaded from a file in one of supported network description languages (e.g. Pajek, GML, JSON), (b) it can be built in the Python code by adding nodes and links manually, (c) it can be constructed from a data structure such as the adjacency matrix or adjacency list.

2.2.3. Interactive network visualisation

One of the strongest features of BEFANA is its interactive visualisation of hierarchical networks. In ecosystem network analysis visualisation is an essential tool for understanding the structure of the network, the interrelations between nodes of interest and for hypothesis formation. BEFANA implements interactive visualisation of ecological networks and responds to user events such as zoom and drag. A network can be visualised either inline in the notebook or in a new browser tab. Inline visualisation follows the computational notebook paradigm and displays a preview of the network and few customization options in a notebook cell while the advanced visualisation in a new tab offers a multitude of options such as colours, shapes, sizes, fonts, layouts and their optimization parameters etc. BEFANA's visualisation also enables modelling the relationship between network topology and network dynamics, which has implications in understanding stability and resilience of trophic networks (Dale 2017). The graphical user interface implements network editing which allows for network modification and analysis cycles. Users can remove certain nodes or edges of a soil food web and then visualise the modified network to assess the extent to which the investigated ecosystem will become destabilised (C.L. Murall, McCann, and Bauch 2012), for instance by functional biodiversity decline (Conti et al., 2020). This loss of nodes will impact the subsequent calculations and will enable us to forecast various scenarios of biodiversity loss.

2.2.4. Network analysis

BEFANA offers network analysis methods that characterise the form of the ecological network and determine how the functioning of the network affects its structure (Dale 2017). While species interactions are central, graph theory can also reveal higher level (guild to guild) interactions, such as how topology of such networks also interacts with the functioning of the ecological system.

Various centrality measures are implemented that give insight on the nodes/edges topological position and role within the network. The out/in-degree ratio (van den Brink and Rusinowska 2021) can be calculated, which assigns to every node in a directed graph its out-degree divided by its in-degree (to avoid dividing by zero, 1 is added to both the out- as well as in-degree of every node). It can be used for ranking nodes in a directed graph. This measure is similar to the well-known Copeland score for ranked voting but can give different results in certain

---

[2] MyBinder: https://mybinder.org/v2/gh/MartinMarzi/BEFANA/HEAD GitHub: https://github.com/MartinMarzi/BEFANA

situations.

Compartment and motif analysis is available that can reveal common repeated structural elements (e.g. transitive triangle motifs) and how clearly separable such compartments are. Strongly connected components can be identified and, in order to discover meaningful subgraphs, clustering within the network can be performed. The results of clustering (node partitioning) can also be used in BEFANA's network visualisation (see Fig. 2).

An important advantage of using BEFANA on ecological networks is its ability to identify the existence and strengths of indirect interactions that are derived from the existing trophic network. BEFANA implements the discovery and visualisation of two types of derived networks: (1) with shared consumer, and (2) with shared resource.

2.2.5. *Network enrichment and modelling with experimental data*

BEFANA allows the user to embed experimental or field data on the network structure which can provide insight on the dynamics of the studied ecosystem. The network data structure allows for an unlimited number of attributes to be assigned to nodes and links. Attributes can also be used in computations provided by the NetworkX library (where applicable) simply by providing the name of the relevant attribute.

BEFANA also implements quantitative comparison of ranked lists using the rank-biased overlap (RBO) measure (Webber et al., 2010) which measures the similarity of potentially infinite ranked lists. It has one parameter (p) which determines how top-weighted the RBO measure is: the smaller p, the more it is top-weighted. RBO is useful when comparing lists of elements (nodes, trophic groups, edges, etc.) ranked according to the value of a selected attribute, such as biomass, for example.

2.2.6. *Predictive modelling with machine learning*

To employ network data in predictive and descriptive machine learning tasks (e.g. classification, link prediction, clustering, graph evolution) the data (nodes, edges) has to be represented as numerical vectors or node embeddings. BEFANA offers node2vec which is one of the recent machine learning algorithms that can learn a mapping from nodes to a dense feature space while maximising the likelihood of preserving network neighbourhoods of nodes (Grover and Leskovec 2016).

2.3. *Launching BEFANA*

The easiest and recommended way to work with BEFANA is to use the public Binder link.[3] Local deployment using Docker or a Python virtual environment is recommended for advanced users. Moreover, BEFANA includes already computed results of the presented soil food web use case for each notebook. All BEFANA notebooks ensure that the required libraries are installed prior to executing any code.

3. Tool demonstration through soil food-web analysis

In this section we demonstrate the application of BEFANA on a case of a soil food web data where we quantify and analyse the hierarchical characteristics of the trophic network at hand. A hierarchical directed graph has been constructed, representing the detrital soil food web using one regional metaweb of all possible literature-based multitrophic links from a local pool of food webs (Mulder et al. 2005; Sechi et al., 2015). The complete analysis consists of the following steps.

3.1. *Loading the data*

The input data is loaded as a csv file with headers and a custom separator. Two csv files are loaded: (1) an interaction network (soil food web) in the form of an adjacency matrix, and (2) trophic levels of the trophic groups.

3.2. *Network construction*

The network is constructed from the adjacency matrix of the metaweb, encompassing three local plots, hereafter called A, B, and C (Sechi et al. 2015) and the hierarchical structure is obtained from the uploaded trophic levels. No editing on the network has been performed.

3.3. *Network visualisation and exploration*

Next, the constructed network is visualised as shown in Fig. 2. Trophic levels define the layers of the network while the colouring of nodes provides a visualisation of their functional roles. Nodes represent trophic groups and their links to the resources: for example, plant roots are present at the bottom and provide food for plant-feeding nematodes, generalist mites, omnivore nematodes, and a number of other trophic groups. In this way, omnivores and top predators that consume from more than one node and/or more than one trophic level are easily recognizable. Colouring can be done either at functional level (e.g. according to an in-degree clustering rule as presented in the Network analysis section) or at taxonomic level (for instance, different colours at the Kingdom level). The numbers prefixing node names are functional codes for trophic nodes: the first digit provides information on the dominant feeding strategy while the second digit provides cladistic information.

3.4. *Network analysis*

In the next step, various network analysis methods are applied that help to characterise the soil food web and help determine how the functioning of the network affects its structure.

3.4.1. *Network characteristics*

In the first stage of network analysis keystone trophic groups and connections that have a disproportionate effect on the network function are identified through the combination of various centrality measures that give insight on the nodes/edges topological position and role within the network in a similar way to Cheddar (Hudson et al. 2013). Each measure provides different insights on trophic node's properties or position in the ecological network. Next, the out/in-degree ratio (van den Brink and Rusinowska 2021) has been calculated and all nodes have been ranked. Interestingly, in our soil food web structure one node from the second trophic level (bacterivore nematodes) is ranked the same as one of the primary resources (fungi), and higher than other second level nodes.

3.4.2. *Compartment and motif analysis*

In the next stage, the soil food web is analysed for common repeated structural elements and how clearly separable such compartments are. Pairs of nodes with reciprocal link are identified, which also includes cannibalistic relationships. For example, our soil food web contains 9 reciprocal trophic interactions (interspecific loops) of which 5 are cannibalistic (intraspecific loops), mostly occurring at higher trophic levels.

All transitive triangle motifs are also discovered, and an example is visualised in Fig. 3. Transitivity depends on directional interactions and is related to the consistency of relationships amongst nodes (Dale 2017). Thus, transitive triangles hint on the stability of the trophic network as they demonstrate the presence of more stable combinations of trophic interactions where a trophic group has more than one resource. Our soil metaweb contains 122 triangles describing transitive relationships.

Strongly connected components can give an indication about the energy flow within the network. Interestingly, our soil food web contains exactly one strongly-connected component spanning three nodes, which is shown in Fig. 3. Moreover, this component (embedding only

---

[3] MyBinder: https://mybinder.org/v2/gh/MartinMarzi/BEFANA/HEAD

Fig. 2. A hierarchical visualisation of the soil metaweb in BEFANA.

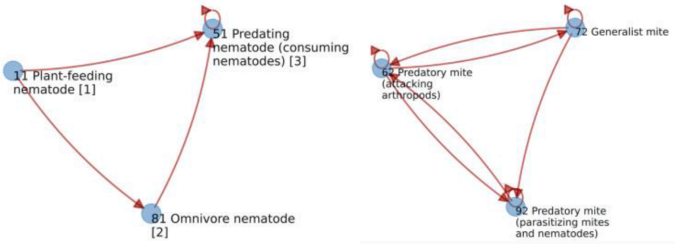


Fig. 3. On the left is an example of a transitive triangle spanning three trophic levels (the number in square brackets denotes trophic level). On the right is the one and only strongly-connected component.

groups from trophic level 4) also contains 8 out of 9 possible links which implies a very high resistance to biodiversity loss. As such, it contains a cycle spanning all three nodes, two cycles spanning two nodes (reciprocal links) and three loops (all three trophic groups are cannibalists). One of the trophic groups is a generalist consumer while the other two are predators.

Finally, clustering within the network is performed. For example, indegree based node clustering into three clusters where indegree is either 0, 1, or greater than 1 yields clusters as shown in Fig. 2. Such clustering separates trophic groups with energy sources from outside of the network, highly vulnerable trophic groups having only one source of energy and less vulnerable having more than one source of energy. With the exception of substrate ingesting earthworms, bacterivore enchytraeid, and bacterivore amoebae all trophic groups on the first level are vulnerable, in contrast to trophic groups on the second and third trophic level (the average indegree is 8.2).

3.4.3. Derived networks

In this step the discovery and visualisation of two types of derived networks is carried out: (1) with shared consumer, and (2) with shared resource. In the analysed soil food web (1) was very connected with 83% of nodes linking to over 60% of all the nodes. In (2), on the other hand, centrality trends in the derived soil food network were clearer: only three nodes had a degree centrality over 60% (omnivore nematode, generalist mite, and omnivore mite).

3.5. Modelling with experimental data

At this stage field data such as aggregated biomass, mass and abundance are computed per trophic group and added to the network as node attributes. This allows for easy computation of the total biomass for each of the trophic levels. Next, rankings of trophic groups by abundance for each of the trophic levels are calculcated and plotted. After that, a ranked list of trophic groups by biomass values for all three plots is produced. Then, this ranked list is compared using the rank-biased overlap (RBO) measure. Regardless of the value of the top-weightedness parameter *p*, plot A is always the most similar to plot C. The least similar are plots A and B if RBO is top-weighted (low p), or plots B and C, if the tails of the biomass value lists are also considered in the comparison (high p). The RBO-based similarity of all three plots using different values of p is shown in Fig. 4. Other trophic groups rank differently according to the plot, and two are not even occurring in all plots. As shown in Fig. 5, earthworms and the three main resources for the plant-driven pathway, the bacterial-driven pathway and the fungal-driven pathway exhibit the highest biomass values in each of the three plots of the investigated metaweb. It is worth mentioning that all groups in common in all separate food webs share a very low coefficient of variation (CV) (less than 25%), in contrast to the group ranked as 21st, that although occupying the same rank according to its biomass values shares a CV higher than 100%.

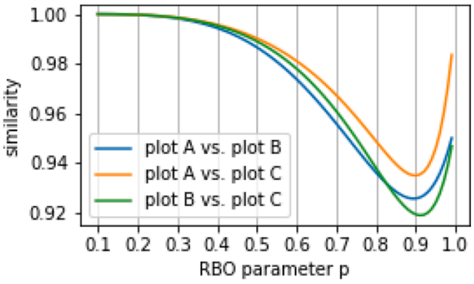
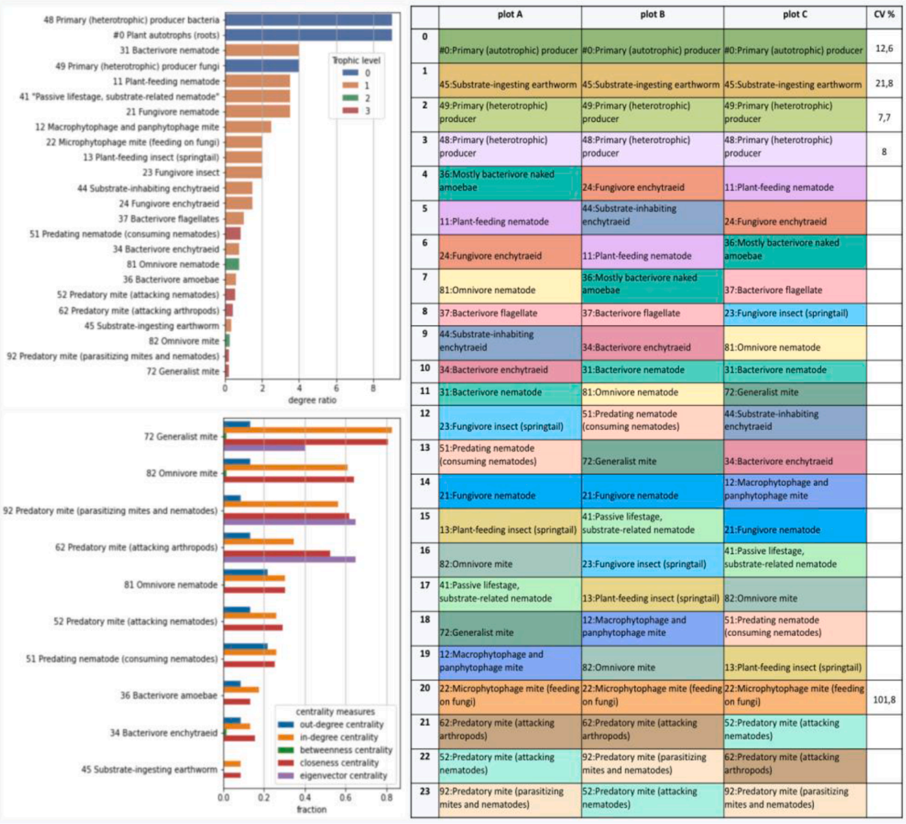


Fig. 4. Similarity of ranked lists of biomass values for local plots A, B, and C using the RBO measure and different values of its top-weightedness parameter *p*.



	plot A	plot B	plot C	CV %
0	#0:Primary (autotrophic) producer	#0:Primary (autotrophic) producer	#0:Primary (autotrophic) producer	12,6
1	45:Substrate-ingesting earthworm	45:Substrate-ingesting earthworm	45:Substrate-ingesting earthworm	21,8
2	49:Primary (heterotrophic) producer	49:Primary (heterotrophic) producer	49:Primary (heterotrophic) producer	7,7
3	48:Primary (heterotrophic) producer	48:Primary (heterotrophic) producer	48:Primary (heterotrophic) producer	8
4	36:Mostly bacterivore naked amoebae	24:Fungivore enchytraeid	11:Plant-feeding nematode	
5	11:Plant-feeding nematode	44:Substrate-inhabiting enchytraeid	24:Fungivore enchytraeid	
6	24:Fungivore enchytraeid	11:Plant-feeding nematode	36:Mostly bacterivore naked amoebae	
7	81:Omnivore nematode	36:Mostly bacterivore naked amoebae	37:Bacterivore flagellate	
8	37:Bacterivore flagellate	37:Bacterivore flagellate	23:Fungivore insect (springtail)	
9	44:Substrate-inhabiting enchytraeid	34:Bacterivore enchytraeid	81:Omnivore nematode	
10	34:Bacterivore enchytraeid	31:Bacterivore nematode	31:Bacterivore nematode	
11	31:Bacterivore nematode	81:Omnivore nematode	72:Generalist mite	
12	23:Fungivore insect (springtail)	51:Predating nematode (consuming nematodes)	44:Substrate-inhabiting enchytraeid	
13	51:Predating nematode (consuming nematodes)	72:Generalist mite	34:Bacterivore enchytraeid	
14	21:Fungivore nematode	21:Fungivore nematode	12:Macrophytophage and panphytophage mite	
15	13:Plant-feeding insect (springtail)	41:Passive lifestage, substrate-related nematode	21:Fungivore nematode	
16	82:Omnivore mite	23:Fungivore insect (springtail)	41:Passive lifestage, substrate-related nematode	
17	41:Passive lifestage, substrate-related nematode	13:Plant-feeding insect (springtail)	82:Omnivore mite	
18	72:Generalist mite	12:Macrophytophage and panphytophage mite	51:Predating nematode (consuming nematodes)	
19	12:Macrophytophage and panphytophage mite	82:Omnivore mite	13:Plant-feeding insect (springtail)	
20	22:Microphytophage mite (feeding on fungi)	22:Microphytophage mite (feeding on fungi)	22:Microphytophage mite (feeding on fungi)	101,8
21	62:Predatory mite (attacking arthropods)	62:Predatory mite (attacking arthropods)	52:Predatory mite (attacking nematodes)	
22	52:Predatory mite (attacking nematodes)	92:Predatory mite (parasitizing mites and nematodes)	62:Predatory mite (attacking arthropods)	
23	92:Predatory mite (parasitizing mites and nematodes)	52:Predatory mite (attacking nematodes)	92:Predatory mite (parasitizing mites and nematodes)	

Fig. 5. The top-left figure shows the degree ratio according to in-degree centrality. The bottom-left figure shows a selection of centrality measures of top-10 nodes. On the right the ranked biomass averages are shown for each separate local plot (A, B and C). The Coefficient of Variation has been computed only for the biomass averages of any trophic group ranked at the same position across the metaweb, five trophic groups in our case. Each colour (plant: green, fungi: brown, bacteria: blue) resembles the pathway, with yellow and reddish for higher trophic level groups (omnivores and predators).

3.6. *Representation learning*

In the last stage of the analysis, BEFANA's representation learning functionality is applied on the derived network as described above. The results demonstrate that the topological information is indeed encoded in node embeddings, and the dimensionality reduction followed by the visualisation in 1D space gives a clearer indication of node similarity. As can be seen in Fig. 6, four trophic groups grouped on the far left are the only nodes at trophic level 1 sharing fungi as resource. Bacterivores, on the other hand, are present and mostly grouped on the right side of the visualization.

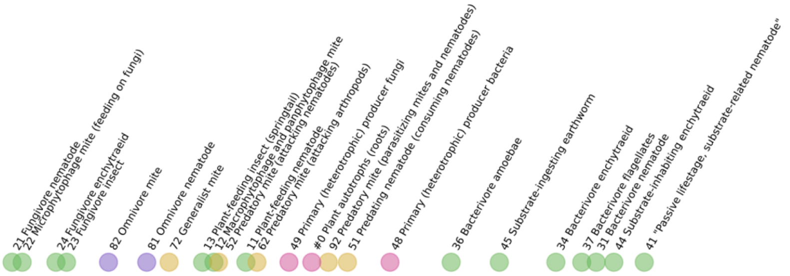


Fig. 6. A visualisation of node embeddings in 1D where the distance between the trophic groups is indicative about node similarity in the derived network. Colours represent trophic levels.

4. Conclusion and future development

The paper presentsed BEFANA, a free and open-source software tool for ecological network visualisation and analysis. It allows ecologists to apply graph theory and machine learning algorithms on ecological networks that can help quantifying and testing hypotheses about how a network's topology affects its dynamics and how the functioning of the network in turn affects its structure. In the future work, we plan to upgrade BEFANA regularly with new features, implementations and datasets. A particular focus will be on machine learning methods optimised for ecological networks. Experimental data is important in ecological analysis and to make use of them machine learning algorithms need to incorporate node and/or edge attributes. While the hierarchical structure of our soil food web makes it unsuitable for representation learning methods, we envision that node embeddings for other types of interaction networks could well be used to predict links when simulating ecological network disruptions.

Data availability

Dryad data: https://doi.org/10.5061/dryad.t5347.

CRediT authorship contribution statement

Martin Marzidovšek: Conceptualization, Formal analysis, Software, Visualization, Writing – original draft, Writing – review & editing. **Vid Podpečan:** Conceptualization, Formal analysis, Software, Visualization, Supervision, Writing – original draft, Writing – review & editing. **Erminia Conti:** Data curation, Methodology, Resources, Validation, Writing – original draft. **Marko Debeljak:** Conceptualization, Formal analysis, Investigation, Project administration. **Christian Mulder:** Conceptualization, Data curation, Investigation, Methodology, Resources, Supervision, Writing – review & editing.

Declaration of Competing Interest

Nothing to declare. The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Martin Marzidovsek reports financial support was provided by Jozef Stefan Institute.

Data Availability

The data has been shared: https://doi.org/10.5061/dryad.t5347

Acknowledgements

This work has been supported by COST action 18237 and project H2020 project EMBEDDIA (#825153). We thank Prof. Dr. Nada Lavrač for her kind support.

References

Bascompte, J., Jordano, P., 2007. Plant-animal mutualistic networks: the architecture of biodiversity. Annu. Rev. Ecol. Evol. Syst. 38 (1), 567–593.

Brink, R.van den, Rusinowska, A., 2021. The degree ratio ranking method for directed graphs. Eur. J. Oper. Res. 288 (2), 563–575.

Cohen, J.E., Jonsson, T., Carpenter, S.R., 2003. Ecological community description using the food web, species abundance, and body size. Proc. Nat. Acad. Sci. 100 (4), 1781–1786.

Cohen, J.E., Schittler, D.N., Raffaelli, D.G., Reuman, D.C., 2009. Food webs are more than the sum of their Tritrophic parts. Proc. Nat. Acad. Sci. 106 (52), 22335–22340.

Conti, E., Di Mauro, L.S., Pluchino, A., Mulder, C., 2020. Testing for Top-down Cascading Effects in a Biomass-Driven Ecological Network of Soil Invertebrates. Ecol. Evo.l 10 (14), 7062–7072.

Dale, M.R.T., 2017. Applying Graph Theory in Ecological Research. Cambridge University Press.

Fitter, A.H., 2005. Darkness visible: reflections on underground ecology. J. Ecol. 93 (2), 231–243.

Grover, A., Leskovec, J., 2016. node2vec: Scalable Feature Learning for Networks. In: In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, pp. 855–864. KDD'16.

Hoogen, J.van den, Geisen, S., et al., 2020. A global database of soil nematode abundance and functional group composition. Sci. Data 7 (1), 103.

Hudson, L.N., et al., 2013. Cheddar: analysis and visualisation of ecological communities in R. Methods in Ecol. Evolution 4 (1), 99–104.

Hunt, H.W., Wall, D.H., 2002. Modelling the effects of loss of soil biodiversity on ecosystem function. Glob. Chang. Biol. 8 (1), 33–50.

Ings, T.C., et al., 2009. Ecological networks–beyond food webs. J. Anim. Ecol. 78 (1), 253–269.

Kondoh, M., Kato, S., Sakato, Y., 2010. Food webs are built up with nested subwebs. Ecology 91 (11), 3123–3130.

McKinney, W., 2010. Data structures for statistical computing in python. In: Proceedings of the 9th Python in Science Conference. https://doi.org/10.25080/Majora-92bf1922-00a.

Mulder, C., Elser, J.J., 2009. Soil acidity, ecological stoichiometry and allometric scaling in grassland food webs. Glob. Chang. Biol. 15 (11), 2730–2738.

Mulder, C., Cohen, J.E., Setälä, H., Bloem, J., Breure, A.M., 2005. Bacterial traits, organism mass, and numerical abundance in the detrital soil food web of dutch agricultural grasslands. Ecol. Lett. 8 (1), 80–90.

Murall, C.L., McCann, K.S., Bauch, C.T., 2012. Food Webs in the Human Body: Linking Ecological Theory to Viral Dynamics. PLoS ONE 7 (11), e48812.

Perez, F., Granger, B.E., 2007. IPython: a system for interactive scientific computing. Comput. Sci. Eng. https://doi.org/10.1109/mcse.2007.53.

Phillips, H.R.P., et al., 2021. Global data on earthworm abundance, biomass, diversity and corresponding environmental properties. Sci. Data 8 (1), 136.

Project Jupyter, et al., 2018. Binder 2.0 - reproducible, interactive, sharable environments for science at scale. In: Akici, F., Lippa, D., Niederhut, D., Pacer, M. (Eds.), Proceedings of the 17th Python in Science Conference.

Reuman, D.C., Cohen, J.E., 2004. Trophic links' length and slope in the tuesday lake food web with species' body mass and numerical abundance. J. Anim. Ecol. 73 (5), 852–866.

Sechi, V., Brussaard, L., De Goede, R.G.M., Rutgers, M., Mulder, C., 2015. Choice of resolution by functional trait or taxonomy affects allometric scaling in soil food webs. Am. Naturalist 185 (1), 142–149.

Webber, W., Moffat, A., Zobel, J., 2010. A similarity measure for indefinite rankings. ACM Trans. Inf. Syst. Security 28 (20), 1–38.

Chapter 6

Conclusions

The dissertation presented how ML when coupled with explainability methods can accelerate and improve the quality of research in the life sciences. This field of science is becoming increasingly important as its sub-disciplines such as ecology, agriculture and biology play a crucial role in understanding the complex threats to the ecosystem and fostering scientific discoveries on which our society's transition to more sustainable patterns of existence depends. The dissertation explored insights into the dynamics of modelled ecosystems that are critical to our food security and our understanding of the underlying factors related to global challenges such as food chain disruption, biodiversity loss and climate change.

While advances in observational technology have facilitated the collection of large and heterogeneous data, the challenges of interpreting and extracting knowledge from these data are increasing, which is why ML methods are becoming so important to advance research in life science disciplines. The dissertation focuses on ecosystem monitoring, ecological modelling, prediction and inference based on ML.

It has been argued that in the application of ML methods in ecology and other sub-disciplines of the life sciences, a balance must be achieved between satisfactory predictive performance on the one hand and sufficient transparency and understanding of the inner workings of the models on the other. Indeed, prediction and research in areas such as life sciences carry considerable weight for policy and decision making, public health and the environment. ML offers advantages for the life sciences, especially for modelling more complex systems with non-linear relationships. However, it has been shown that such models can be difficult to interpret. This limits their usefulness in scientific settings, where the ultimate goal is to understand the data generation processes underlying the patterns that ML reveals. Because of this problem, it has been argued that it is sensible to use interpretable ML models as much as possible and/or to use explainability methods to address the black box nature of more complex ML models by helping to understand their inner workings and decisions.

The dissertation also shows that the successful application of ML techniques in the life sciences requires a collaborative approach between ML and life science experts. The presented research on the application of ML in ecology has shown that it is necessary to integrate ecological domain knowledge into the developed ML methods and to draw knowledge and understanding from the trained models. Most of the work presented here is the result of such cross-domain collaboration.

The overarching goal of this dissertation is to advance the use of xAI and network science in selected sub-disciplines of the life sciences, particularly ecology. This is done by developing ML-based solutions for automated ecological monitoring, pattern recognition, predictive analytics, network analysis and representation learning on graphs. The dissertation emphasises the explainability of the underlying ML methods, which is crucial for the life sciences due to their impact on the real world. In particular, ML methods are developed that can accelerate different phases of traditional ecological research and promote research in other life sciences that can provide answers and solutions for dealing with the escalating impacts of climate change and human-induced stressors on biodiversity and ecosystems. The specific contributions of ML solutions to the life sciences, which form the main part of this dissertation, are summarised below.

- In the first contribution, an explainable ML approach is developed to predict the toxicity of mussels in the Gulf of Trieste caused by HABs. We created and analysed a 28-year dataset containing records of toxic phytoplankton in mussel farming areas and toxin concentrations in mussels as well as other environmental data. We found that the right data pre-processing steps are crucial to overcome the specific challenges of the consolidated dataset. Model training and performance evaluation show that RF performed best among the models obtained.

  Explainability methods are used to gain insights into the decisions of the model and thus provide more transparency, which is needed in real-world scenarios such as EWS. The applied methods also enable a better understanding of the dynamics of the marine ecosystem influenced by environmental and climatic pressures. The solution developed can be a cost-effective means of implementing measures to minimise health risks and social and economic damage.

- The second contribution is the application of usupervised ML in agriculture. The aim was to identify the barriers and incentives for the use of DSS for IPM. Cluster analysis provided insights into the variability of survey participants' responses regarding their experiences and expectations when using DSS in IPM. Both farmers and farm advisors showed a tendency to form three groups. Outliers were found among both farmers and farm advisors that could not be assigned to any of the detected clusters by the HDBSCAN algorithm. Explanation by simplification and visual explanation then identified the most common barriers and incentives to access and adoption of DSS by farmers and farm advisors that characterise each of the identified clusters. These findings allow the experts to direct future research towards finding suitable solutions to overcome these barriers.

- The next contribution of this dissertation is a developed computer vision system to identify and quantify phytoplankton species, estimate cell size and calculate biovolume based on cell morphology. Such a system can be used to study how climatic changes affect phytoplankton physiology and community structure, thereby altering marine ecosystems and their functions. The method was developed using transfer learning, in which generic, pre-trained models were fine-tuned with microscope images from the Adriatic Sea. Compared to manual analysis, this method processes more samples and provides a more ecologically relevant assessment with high confidence due to the visual explanations.

- The final contribution is the development of BEFANA, a special software tool that enables ecologists to apply network analyses, interactive visualisation and ML algorithms to ecological networks. It helps them to quantify and test hypotheses about how the topology of a network influences its dynamics and how the functioning of the network in turn influences its structure. It is a free and open-source software tool designed to make it easier for ecologists to access common analytical tasks in ecological networks.

  BEFANA provides methods for data loading and pre-processing, network analysis and interactive visualisation, and modelling with experimental data. To utilise the network data for predictive and descriptive ML tasks, the data (nodes, edges) can be represented as numerical vectors or node embeddings using the node2vec ML algorithm. BEFANA enables the enrichment and modelling of networks with experimental data. The user can embed experimental or field data into the network structure, which can provide information about the dynamics of the ecosystem under investigation. BEFANA was evaluated by applying it to a soil food web of an agricultural grassland.

Five hypotheses were put forward in the dissertation, which were dealt with in the above-mentioned contributions. The result of the dissertation is that all five hypotheses are fully confirmed by the results, as shown in the corresponding sections above. The planned further work in connection with the individual contributions was presented in the corresponding chapters and in the publications above.

In line with the principles of open, reproducible and transparent science, the presented work and results, including the software code and data, are openly accessible (see papers for code and data repositories), while the scientific publications are all in open access.

References

Arik, S. Ö., & Pfister, T. (2021). TabNet: Attentive interpretable tabular learning. *AAAI*, *35*(8), 6679–6687.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, *58*, 82–115.

Bau, D., Zhou, B., Khosla, A., Oliva, A., & Torralba, A. (2017). Network dissection: Quantifying interpretability of deep visual representations. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6541–6549.

Borowiec, M. L., Dikow, R. B., Frandsen, P. B., McKeeken, A., Valentini, G., & White, A. E. (2022). Deep learning as a tool for ecology and evolution. *Methods Ecol. Evol.*, *13*(8), 1640–1660.

Breiman, L. (2001). Random forests. *Mach. Learn.*, *45*(1), 5–32.

Brown, A. R., Lilley, M. K., Shutler, J., Widdicombe, C., Rooks, P., McEvoy, A., Torres, R., Artioli, Y., Rawle, G., Homyard, J., et al. (2022). Harmful algal blooms and their impacts on shellfish mariculture follow regionally distinct patterns of water circulation in the western english channel during the 2018 heatwave. *Harmful Algae*, *111*, 102166.

Bussonnier, M., Jupyter, P., Forde, J., Freeman, J., Granger, B., Head, T., Holdgraf, C., Kelley, K., Nalvarte, G., Osheroff, A., Pacer, M., Panda, Y., Perez, F., Ragan-Kelley, B., & Willing, C. (2018). Binder 2.0 - reproducible, interactive, sharable environments for science at scale. *F. Akici, D. Lippa, D. Niederhut,M. Pacer, eds.*

Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-Based clustering based on hierarchical density estimates. *Advances in Knowledge Discovery and Data Mining*, 160–172.

Caret, K. M. (2018). Classification and regression training. R package version 6.0-76; 2017.

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, *8*(8), 832.

Christin, S., Hervet, É., & Lecomte, N. (2019). Applications for deep learning in ecology. *Methods Ecol. Evol.*, *10*(10), 1632–1644.

Cios, K. J., Kurgan, L. A., & Reformat, M. (2007). Machine learning in the life sciences (vol 26, pg 14, 2007). *IEEE Eng. Med. Biol. Mag.*, *26*(3), 9–9.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Mach. Learn.*, *20*(3), 273–297.

Dale, M. R. (2017). *Applying graph theory in ecological research*. Cambridge University Press.

Debeljak, M., & Džeroski, S. (2011). Decision trees in ecological modelling. In F. Jopp, H. Reuter, & B. Breckling (Eds.), *Modelling complex ecological dynamics: An introduction into ecological modelling for students, teachers & scientists* (pp. 197–209). Springer Berlin Heidelberg.

Desjardins-Proulx, P., Laigle, I., Poisot, T., & Gravel, D. (2017). Ecological interactions and the netflix problem. *PeerJ*, *5*, e3644.

Domingos, P. (2012). A few useful things to know about machine learning. *Commun. ACM*, *55*(10), 78–87.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608.*

Dzeroski, S., & Todorovski, L. (2007, August). *Computational discovery of scientific knowledge: Introduction, techniques, and applications in environmental and life sciences.* Springer.

Eikelboom, J. A. J., Wind, J., van de Ven, E., Kenana, L. M., Schroder, B., de Knegt, H. J., van Langevelde, F., & Prins, H. H. T. (2019). Improving the precision and accuracy of animal population estimates with aerial image object detection. *Methods Ecol. Evol.*, *10*(11), 1875–1887.

Elith, J., H. Graham, C., P. Anderson, R., Dudík, M., Ferrier, S., Guisan, A., J. Hijmans, R., Huettmann, F., R. Leathwick, J., Lehmann, A., Li, J., G. Lohmann, L., A. Loiselle, B., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., McC. M. Overton, J., Townsend Peterson, A., … E. Zimmermann, N. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, *29*(2), 129–151.

Farley, S. S., Dawson, A., Goring, S. J., & Williams, J. W. (2018). Situating ecology as a Big-Data science: Current advances, challenges, and solutions. *Bioscience*, *68*(8), 563–576.

Ferreira, A. C., Silva, L. R., Renna, F., Brandl, H. B., Renoult, J. P., Farine, D. R., Covas, R., & Doutrelant, C. (2020). Deep learning-based methods for individual recognition in small birds. *Methods Ecol. Evol.*, *11*(9), 1072–1085.

Gobeyn, S., Mouton, A. M., Cord, A. F., Kaim, A., Volk, M., & Goethals, P. L. M. (2019). Evolutionary algorithms for species distribution modelling: A review in the context of machine learning. *Ecol. Modell.*, *392*, 179–195.

Grover, A., & Leskovec, J. (2016). Node2vec: Scalable feature learning for networks. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 855–864. https://doi.org/10.1145/2939672.2939754

Guo, Q., Jin, S., Li, M., Yang, Q., Xu, K., Ju, Y., Zhang, J., Xuan, J., Liu, J., Su, Y., Xu, Q., & Liu, Y. (2020). Application of deep learning in ecological resource research: Theories, methods, and challenges. *Sci. China Earth Sci.*

Hagberg, A., Swart, P., & S Chult, D. (2008). *Exploring network structure, dynamics, and function using networkx* (tech. rep.). Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

Hastie, T., Friedman, J., & Tibshirani, R. (2009). *The elements of statistical learning.* Springer New York.

Heil, B. J., Hoffman, M. M., Markowetz, F., Lee, S.-I., Greene, C. S., & Hicks, S. C. (2021). Reproducibility standards for machine learning in the life sciences. *Nat. Methods*, *18*(10), 1132–1135.

Holzinger, A., Keiblinger, K., Holub, P., Zatloukal, K., & Müller, H. (2023). AI for life: Trends in artificial intelligence for biotechnology. *N. Biotechnol.*, *74*, 16–24.

Hunter. (2007). Matplotlib: A 2D graphics environment. *9*, 90–95.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260.

Karlson, B., Andersen, P., Arneborg, L., Cembella, A., Eikrem, W., John, U., West, J. J., Klemm, K., Kobos, J., Lehtinen, S., Lundholm, N., Mazur-Marzec, H., Naustvoll, L., Poelman, M., Provoost, P., De Rijcke, M., & Suikkanen, S. (2021). Harmful algal blooms and their effects in coastal seas of northern europe. *Harmful Algae*, *102*, 101989.

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B. E., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J. B., Grout, J., Corlay, S., et al. (2016). Jupyter notebooks-a publishing format for reproducible computational workflows. *Elpub*, *2016*, 87–90.

Kwok, R. (2019). AI empowers conservation biology. *Nature*, *567*(7746), 133–134.

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521(7553).*

Loizides, F., & Schmidt, B. (2016, May). *Positioning and power in academic publishing: Players, agents and agendas: Proceedings of the 20th international conference on electronic publishing.* IOS Press.

Lucas, T. C. D. (2020). A translucent box: Interpretable machine learning in ecology. *Ecol. Monogr.*, *90*(4).

Mac Aodha, O., Gibb, R., Barlow, K. E., et al. (2018). Bat detective—deep learning tools for bat acoustic signal detection. *PLoS Comput. Biol.*

Maglianesi, M. A., Blüthgen, N., Böhning-Gaese, K., & Schleuning, M. (2014). Morphological traits determine specialization and resource use in plant–hummingbird networks in the neotropics. *Ecology*, *95*(12), 3325–3334.

Masís, S. (2023). *Interpretable machine learning with python: Build explainable, fair, and robust high-performance models with hands-on, real-world examples.* Packt Publishing Ltd.

McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426.*

McKenzie, C. H., Bates, S. S., Martin, J. L., Haigh, N., Howland, K. L., Lewis, N. I., Locke, A., Peña, A., Poulin, M., Rochon, A., Rourke, W. A., Scarratt, M. G., Starr, M., & Wells, T. (2021). Three decades of canadian marine harmful algal events: Phytoplankton and phycotoxins of concern to human and ecosystem health. *Harmful Algae*, *102*, 101852.

McKinney, W., et al. (2010). Data structures for statistical computing in python. *Proceedings of the 9th Python in Science Conference*, *445*, 51–56.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, *267*, 1–38.

Mitchel, T. (1997). Machine learning, McGraw-Hill education (ISE editions).

Molnar, C. (2022). *Interpretable machine learning: A guide for making black box models explainable* (2nd ed.). https://christophm.github.io/interpretable-ml-book

Moulavi, D., Jaskowiak, P. A., Campello, R. J. G. B., Zimek, A., & Sander, J. (2014, April). Density-Based clustering validation. In *Proceedings of the 2014 SIAM international conference on data mining (SDM)* (pp. 839–847). Society for Industrial; Applied Mathematics.

Olah, C., Mordvintsev, A., & Schubert, L. (2017). Feature visualization. *Distill*, *2*(11).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2012). Scikit-learn: Machine learning in python, 2825–2830.

Perez, F., & Granger, B. E. (2007). IPython: A system for interactive scientific computing. *Computing in Science Engineering*, *9*(3), 21–29.

Peters, D. P. C., Havstad, K. M., Cushing, J., Tweedie, C., Fuentes, O., & Villanueva-Rosales, N. (2014). Harnessing the power of big data: Infusing the scientific method with machine learning to transform ecology. *Ecosphere*, *5*(6), art67.

Phillips, S. J., & Dudík, M. (2008). Modeling of species distributions with maxent: New extensions and a comprehensive evaluation. *Ecography*, *31*(2), 161–175.

Pichler, M., Boreux, V., Klein, A.-M., Schleuning, M., & Hartig, F. (2020). Machine learning algorithms to infer trait-matching and predict species interactions in ecological networks. *Methods Ecol. Evol.*, *11*(2), 281–293.

Pichler, M., & Hartig, F. (2023). Machine learning and deep learning—a review for ecologists. *Methods Ecol. Evol.*, *14*(4), 994–1016.

Poisot, T., Stouffer, D. B., & Gravel, D. (2015). *Beyond species: Why ecological interaction networks vary through space and time* (No. *3*).

Raghu, M., & Schmidt, E. (2020). A survey of deep learning for scientific discovery, 1–48.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier, 1135–1144.

Ripley, B. D. (2007). *Pattern recognition and neural networks.* Cambridge University Press.

Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A. S., Milojevic-Dupont, N., Jaques, N., Waldman-Brown, A., Luccioni, A., Maharaj, T., Sherwin, E. D., Mukkavilli, S. K., Kording, K. P., Gomes, C., Ng, A. Y., Hassabis, D., Platt, J. C., … Bengio, Y. (2022). Tackling climate change with machine learning. *ACM Computing Surveys*, (Issue 2).

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, *1*(5), 206–215.

Rüping, S. (2006). Learning interpretable models (ph. d. thesis). *University Dortmund.*

Ryo, M., Angelov, B., Mammola, S., Kass, J. M., Benito, B. M., & Hartig, F. (2021). Explainable artificial intelligence enhances the ecological interpretability of blackbox species distribution models. *Ecography*, *44*(2), 199–205.

Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. *Commun. ACM*, *63*(12), 54–63.

Shapley, L. (2020, November). 7. a value for n-person games. contributions to the theory of games II (1953) 307-317. In *Classics in game theory* (pp. 69–79). Princeton University Press.

Strydom, T., Catchen, M. D., Banville, F., Caron, D., Dansereau, G., Desjardins-Proulx, P., Forero-Muñoz, N. R., Higino, G., Mercier, B., Gonzalez, A., Gravel, D., Pollock, L., & Poisot, T. (2021). A roadmap towards predicting species interaction networks (across space and time). *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, *376*(1837), 20210063.

Tabak, M. A., Norouzzadeh, M. S., Wolfson, D. W., Newton, E. J., Boughton, R. K., Ivan, J. S., Odell, E. A., Newkirk, E. S., Conrey, R. Y., Stenglein, J., Iannarilli, F., Erb, J., Brook, R. K., Davis, A. J., Lewis, J., Walsh, D. P., Beasley, J. C., VerCauteren, K. C., Clune, J., & Miller, R. S. (2020). Improving the accessibility and transferability of machine learning algorithms for identification of animals in camera trap images: MLWIC2. *Ecol. Evol.*, *10*(19), 10374–10383.

Thessen, A. E. (2016). Adoption of machine learning techniques in ecology and earth science. *One Ecosystem*, *1*, 1–38.

Thuiller, W., Georges, D., Engler, R., Breiner, F., Georges, M. D., & Thuiller, C. W. (2016). Package 'biomod2'. *Species distribution modeling within an ensemble forecasting framework.*

Tuia, D., Kellenberger, B., Beery, S., Costelloe, B. R., Zuffi, S., Risse, B., Mathis, A., Mathis, M. W., van Langevelde, F., Burghardt, T., Kays, R., Klinck, H., Wikelski, M., Couzin, I. D., van Horn, G., Crofoot, M. C., Stewart, C. V., & Berger-Wolf, T. (2022). Perspectives in machine learning for wildlife conservation. *Nat. Commun.*, *13*(1), 792.

United Nations. (2015). Transforming our world: The 2030 agenda for sustainable development. https://wedocs.unep.org/20.500.11822/9814

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research 9 (11).*

Wäldchen, J., & Mäder, P. (2018). Machine learning for image based species identification. *Methods Ecol. Evol.*, *9*(11), 2216–2225.

Wang, Y., Yao, Q., Kwok, J. T., & Ni, L. M. (2020). Generalizing from a few examples: A survey on few-shot learning. *ACM Comput. Surv.*, *53*(3), 1–34.

Waskom, M. (2021). Seaborn: Statistical data visualization. *J. Open Source Softw.*, *6*(60), 3021.

Weinstein, B. G. (2018). A computer vision for animal ecology. *J. Anim. Ecol.*, *87*(3), 533–545.

Bibliography

Publications Related to the Thesis

Journal Articles

Marinko, J., Ivanovska, A., Marzidovšek, M., Ramsden, M., & Debeljak, M. (2023). Incentives and barriers to adoption of decision support systems in integrated pest management among farmers and farm advisors in europe. *International Journal of Pest Management*, 1–18.

Marzidovšek, M., Francé, J., Podpečan, V., Vadnjal, S., Dolenc, J., & Mozetič, P. (2024). Explainable machine learning for predicting diarrhetic shellfish poisoning events in the adriatic sea using long-term monitoring data. *Harmful Algae*, 102728.

Marzidovšek, M., Mozetič, P., Francé, J., & Podpečan, V. (2024b). Computer vision techniques for morphological analysis and identification of two pseudo-nitzschia species. *Water*, *16*(15), 2160.

Marzidovšek, M., Podpečan, V., Conti, E., Debeljak, M., & Mulder, C. (2022). Befana: A tool for biodiversity-ecosystem functioning assessment by network analysis. *Ecological Modelling*, *471*, 110065.

Conference Papers

Marzidovšek, M., Mozetič, P., Francé, J., & Podpečan, V. (2024a). Analyzing the phytoplankton community structure with computer vision. *Nordic Oikos 2024: abstract book*, 83.

Marzidovšek, M., Podpečan, V., Mozetič, P., Francé, J., Dolenc, J., & Grebenc, S. (2023). Explaining machine learning models for predicting harmful impact of algae blooms on aquaculture. *Machine learning and data analysis in oceanography*, 1.

Other Publications

Conference Papers

Martinc, M., Škrlj, B., Pirkmajer, S., Lavrač, N., Cestnik, B., Marzidovšek, M., & Pollak, S. (2020). Covid-19 therapy target discoverywith context-aware literature mining. *12323*, 109–123.

Marzidovšek, M., & Pollak, S. (2020). A study on reproducibility in computational creativity research: Lessons learned and new prospects, 350–357.

Marzidovšek, M., Ramšak, Ž., Gruden, K., & Podpečan, V. (2023). Link prediction on comprehensive knowledge graph. *15th Meeting of the Slovenian Biochemical Society with International Participation: book of abstracts*, 87.

Biography

Martin Marzidovšek holds a Master's degree in International Affairs from the Graduate Institute of International and Development Studies at the University of Geneva and is currently pursuing a PhD at the Jožef Stefan Postgraduate School, Information and Communication Technologies Programme.

As a researcher at the Jozef Stefan Institute (JSI), his research mainly revolved around the application of machine learning in the life sciences. He had the privilege of working on various projects in the fields of marine biology, molecular and systems biology, ecology and agriculture. He collaborated with experts from the Slovenian National Institute of Biology, the Department of Biology, Geology and Environmental Sciences at the University of Catania, Italy and the California Institute of Technology. During his time at JSI, he was also the exploitation manager for the EU-funded international project EMBEDDIA.

Before joining JSI, he was actively involved in the tech startup ecosystem in Germany and Slovenia. His entrepreneurial background includes co-founding a venture capital funded tech startup that was selected for business accelerators in Berlin and Silicon Valley. Prior to this, he worked in various international sectors, including for-profit and non-profit organisations.

He was a member of the National Assembly of the Republic of Slovenia (2022–2023) and is currently an advisor for AI at the Ministry for digital transformation of the Republic of Slovenia.